ORIGINAL ARTICLE

Reinforcement learning with intrinsic affinity for personalized prosperity management

Charl Maree[1,2] · **Christian W. Omlin[1]**

Abstract

The purpose of applying reinforcement learning (RL) to portfolio management is commonly the maximization of profit. The extrinsic reward function used to learn an optimal strategy typically does not take into account any other preferences or constraints. We have developed a regularization method that ensures that strategies have global intrinsic affinities, i.e., different personalities may have preferences for certain asset classes which may change over time. We capitalize on these intrinsic policy affinities to make our RL model inherently interpretable. We demonstrate how RL agents can be trained to orchestrate such individual policies for particular personality profiles and still achieve high returns.

Keywords AI in banking · Personalized financial services · Explainable AI · Reinforcement learning · Policy regularization · Intrinsic affinity · Robo-advising

JEL Classification C10 · C30 · C32 · C40 · C45 · C50 · C51 · C52 · C53 · C54 · C58 · D10 · D14 · D31 · D53 · D91 · E22 · E37 · G11 · G41

1 Introduction

Effective customer engagement is a prerequisite for modern financial service providers that are adopting advanced methods to increase the level of personalization of their services (Stefanel & Goyal, 2019). Although artificial intelligence (AI) has become a ubiquitous tool in financial technology (Fernández, 2019), research in the field has yet to significantly advance levels of personalization (Maree & Omlin,

✉ Charl Maree
  charl.maree@uia.no

  Christian W. Omlin
  christian.omlin@uia.no

1   Center for Artificial Intelligence Research, University of Agder, Grimstad, Norway

2   Chief Technology Office, Sparebank 1 SR-Bank, Stavanger, Norway

2021). Asset management is an active research topic in AI for finance; however, research opportunities presented by the need for personalized services are usually neglected (Millea, 2021). Whereas personalized investment advice is typically based on questionnaires, we propose a customer profiling from micro-segmentation that is based on spending behavior. Traditionally, customer segmentation has been grounded in demographics that provide only a coarse segmentation (Smith, 1956); it fails to capture nuanced differences between individuals with the potential for undesirable ramifications, e.g. discrimination in credit scoring based on postal code (Barocas & Selbst, 2016). Micro-segmentation, however, provides a more sophisticated classification that can improve the quality of banking services (Mousaeirad, 2020; Apeh et al., 2011).

We develop a personal prosperity manager that invests in a portfolio of asset classes according to individual personality profiles, as manifested by their spending behavior. The result is a hierarchical system of reinforcement learning (RL) agents in which a high-level agent orchestrates the actions of five low-level agents with global intrinsic affinities for certain asset classes. These affinities derive from prototypical personality traits. For instance, personality traits with a higher affinity for risk may, as a general rule, prefer high-volatility asset types.

Explainability and interpretability form the basis for understanding and trust (Barredo Arrieta et al., 2020). They are imperative for critical industries such as finance, but they have not yet been adequately addressed (Ramon et al., 2021; Cao, 2021). We regularize our agents' policies by predefined prior action distributions, thus imprinting characteristic behaviors that make their policies inherently interpretable on three levels: (1) the salient features extracted from customer spending behavior, (2) the affinities of the prototypical agents, and (3) their orchestration to achieve personal investment advice. Our contribution is, therefore, twofold: we demonstrate how RL agents can be made inherently interpretable through their intrinsic affinities, and we exemplify their value through their application in personalized prosperity management.

2 Background and related work

Recurrent neural networks (RNNs) are a class of artificial deep neural networks that are adept at processing temporal inputs. Their nodes maintain a memory of past events and learn representations in the form of activations (Hochreiter & Schmidhuber, 1997). It is established practice to investigate these node activations using the tools provided by the theory of dynamical systems (Ceni et al., 2019; Maheswaranathan et al., 2019). The state space of a RNN refers to the N-dimensional representation of the node activations, where N is the number of nodes in the RNN. For three (or fewer) nodes, their activation can, for example, be visualized in a three (or lower) dimensional plot, where each axis represents one node. This state-space plot is a useful implement for investigating the dynamics that govern the RNN. The theory of dynamical systems introduces the concept of attractors (Milnor, 2004); they are a set of states, or points in the state space, toward which neighboring states converge.

There are two main classes of attractors: fixed attractors, e.g., points, lines, surfaces, or other geometric shapes, and strange attractors that cannot be described as combinations of these, e.g., oscillating, chaotic, etc.

Gladstone et al. (2019) found that spending patterns are a predictor of financial personality. They trained a random forest to predict customer personalities from their classified financial transactions, using a prevalent taxonomy of personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Although they achieved only a modest predictive accuracy, Tovanich et al. (2021) found that spending patterns over time expose salient information that is obscured in non-temporal form; the authors in this study used the same personality model, but added temporal patterns such as variability of the amount, persistence of the category in time, and burstiness—the intermittent changes in frequency of an event. Recurrent neural networks are able to extract this salient information when predicting personality traits from financial transactions (Maree & Omlin, 2021). In Maree and Omlin (2022c), we gained an understanding of these extracted features by interpreting the dynamics of the RNN state space through locating the set of attractors that govern the model. Understanding model behavior is crucial in industries such as personal finance (Ramon et al., 2021). In their study, Ramon et al. (2021) extracted rules from three classes of models—linear regression, logistic regression, and random forests—which not only exposed the spending patterns most indicative of personality traits, but also aided in model improvement.

In RL, agents learn to solve problems by tentation; they maximize the expected rewards resulting from their actions in an environment (Sutton & Barto, 2018). The expected reward R is the sum of discounted rewards for a time horizon controlled by a discount factor γ: $R = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^T r_{t+T}$. The environment is modelled as a Markov decision process (MDP) with sets of states S, actions A, rewards $R(s, a)$, $s \in S, a \in A$, and transition probabilities $P(s'|s, a)$. Deterministic policy gradients (DDPG, Lillicrap et al., 2019) is an algorithm that maximizes the expected reward by learning a state-action value function $Q(s, a)$ and the optimum action for each state $\mu(s)$. For numerical stability, it uses duplicate 'target' models $Q'(s, a)$ and $\mu'(s)$ for which the parameters θ' are updated slowly using the soft-update formula: $\theta' = \tau\theta + (1 - \tau)\theta'$ where τ is normally a small value and θ and θ' refer to the main and target network parameters, respectively. Environments can have complex dynamics that result in sophisticated policies that are opaque to their developers, who may neither understand nor be able to control what these agents learn (Heuillet et al., 2021; García & Fernández, 2015). Intrinsic motivation enables agents to learn behaviors that are detached from the expected rewards of the environment (Aubret et al., 2019). It is a strategy that was developed to address the challenge of exploration in environments with sparse rewards (Andres et al., 2022). One such approach is Kullback-Leibler (KL) policy regularization in which the objective function is regularized by the KL-divergence between the current policy and a predefined prior (Galashov et al., 2019). Policy regularization has been shown to be helpful and never detrimental to convergence (Vieillard et al., 2022). Although most policy regularization methods aim to improve learning performance, they can also control the learning

process and imbue the policy with an intrinsic behavior (Maree & Omlin, 2022a). Here, the DDPG objective function is regularized with a predefined prior action distribution that defines a desirable characteristic:

$$J(\theta) = \mathbb{E}_{s,a \sim \mathcal{D}}[R(s,a)] - \lambda L$$

$$L = \frac{1}{M} \sum_{j=0}^{M} \left[\mathbb{E}_{a \sim \pi_\theta} \left[a_j \right] - (a_j | \pi_0(a)) \right]^2 \qquad (1)$$

$J(\theta)$ is the learning objective as a function of the model parameters θ, $R(s,a)$ is the expected reward for state s and action a as sampled from a replay buffer $\mathcal{D}$, and λ is a scaling hyperparameter for the regularization term L, which is the mean square difference across M number of actions between the current action distribution and the action distribution given a regularization prior π_0. The efficacy of this approach was demonstrated by instilling an inherent characteristic behavior in agents that navigate a grid. These agents learned to either prefer left turns, right turns, or to avoid going straight by taking a zig-zag approach to their destination. In contrast to constrained RL which *avoids* certain states (Miryoosefi et al., 2019), the policy regularization in Maree and Omlin (2022a) *encourages* certain actions irrespective of the state and is a new direction for RL.

Hierarchical reinforcement learning (HRL) decomposes problems into low-level subtasks that are learned by relatively simple agents for the purpose of either improved performance or explainability (Pateria et al., 2021; Levy et al., 2019). Larger problems are solved by choreographing these subtasks through an orchestration agent that learns the high-level dynamics of its environment (Hengst, 2010). To our knowledge, there have been no applications of HRL in finance, and our work is the first. HRL has, however, been used to control a robotic arm: while low-level agents learned simple tasks such as moving forward/backward or picking up/placing down, an orchestration agent learned to retrieve objects on a surface by choreographing these tasks (Marzari et al., 2021; Beyret et al., 2019). The agents were not only efficient at learning, but their policies were more easily interpreted by human experts. Kulkarni et al. (2016) used HRL to train a hierarchical set of agents to play a game. Their low-level agents learned to solve simple tasks such as "pick up a key" or "open a door" while receiving extrinsic rewards from the environment. A high-level agent then orchestrated these sub-tasks and received intrinsic rewards generated by a critic based on whether or not larger objectives were met.

3 Methodology

To facilitate a comprehensive understanding of our work, we give a brief summary of previous work in learning prototypical investment strategies and customer profiling based on spending behavior. We discuss how we trained five low-level RL agents to invest in a set of asset classes according to prototypical personality traits (Maree & Omlin, 2022b), and how we extracted spending-behavioral trajectories from the state space of a RNN that predicts personality from financial transactions (Maree &

Omlin, 2021). We then detail our approach in combining these preliminaries to learn unique and personal compositions of prototypical strategies using hierarchical RL. Finally, we discuss our methodology of learning temporal strategies using several such compositions in a RNN. These temporal strategies eliminate the need to retrain orchestration agents when customers' spending behavior change, or for new customers. We illustrate this process in a flow diagram in Fig. 1.

3.1 Personality-based profiling

We have previously developed a three-node RNN that predicts customer personalities from an input vector of their classified financial transactions (Maree & Omlin, 2021). This input vector consists of six annual time steps, each consisting of 97 transaction classes; the values in each time step add up to one and are the fraction of a customer's annual spending per transaction category. The RNN output
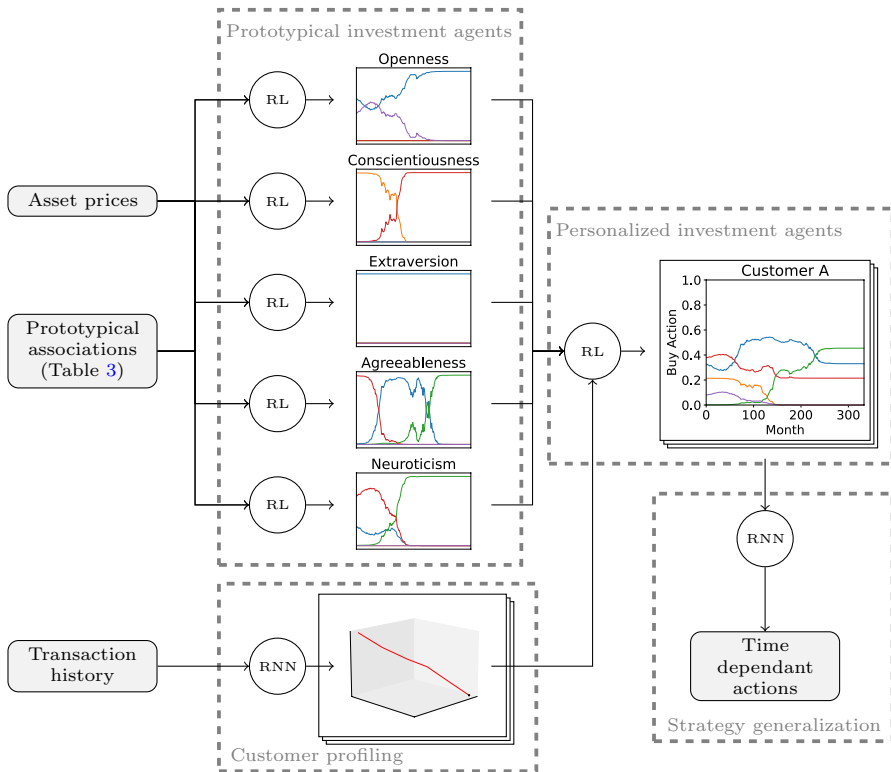

Fig. 1 Information flow diagram illustrating how our system uses financial transactions to generate personalized investment advice. We use hierarchical RL agents with intrinsic affinity to learn unique compositions of prototypical investment strategies that match personal financial preferences. We use many of these compositions to train a RNN to predict a final composition which allows for shifting strategies in time and eliminates the need to retrain an orchestration agent for each unique customer
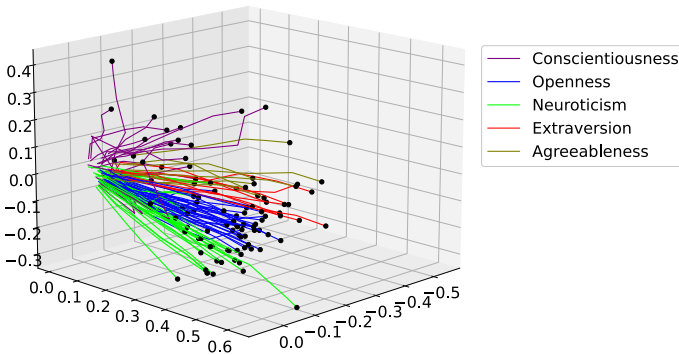
Fig. 2 Clustering behavior of a subset of 100 trajectories in the state space of a RNN. Each trajectory represents a customer's spending behavior in time and is labelled according to the customer's dominant personality trait. Each axis is the activation of one of the three nodes in the RNN

is a five-dimensional personality vector; its values are the degrees of membership in each of five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. We use the feature trajectories from this model's state space—shown in Fig. 2—to represent a customer's spending behavior over time.

Each behavioral trajectory represents an individual customer and is labeled according to their most dominant personality trait: the trait with the greatest value in the personality vector. Linear trajectories indicate consistent spending behavior in time, while trajectories that veer from their initial direction indicate that that customer had changed their spending behavior. This explains why some trajectories seem to behave differently from others of the same color. We refer to Maree and Omlin (2022c) for a detailed discussion about the behavior of these trajectories. These trajectories form clusters in the state space, which separate into sub-clusters along the successive levels of lesser personality traits. This hierarchical clustering provides a means of micro-segmenting customers according to their spending behavior in time. We then explained these behavioral trajectories by reproducing them using a linear regression model, and we interpreted them through locating a number of attractors that govern the dynamics of the state space (Maree & Omlin, 2022c). We located these attractors by mapping the RNN output space into the state space through inverse regression. Using this mapping, and the maximum *reachable* values in the output space, based on the known range of the dimensions in the state space ($[-1, 1]$), we extrapolated the final locations (attractors) of the behavioral trajectories. Formally:

$$\mathcal{O} = \mathcal{D} \cdot \boldsymbol{\omega}_{\text{inv}} - \vec{0} \cdot \boldsymbol{\omega}_{\text{inv}}$$

$$\mathcal{D} = \text{diag}\left\{ \max_{1 \leq i \leq |K|} \mathbf{O}_{i,j}, \ j \in [1..P] \right\}$$

$$\boldsymbol{\omega}_{\text{inv}} = (\mathbf{O}^{\text{T}}\mathbf{O})^{-1} \cdot (\mathbf{O}^{\text{T}}\mathbf{S})$$

$$\mathbf{O} = \mathbf{S} \cdot \boldsymbol{\omega}_{\text{out}}$$

where $\mathcal{O} \in \mathbb{R}^{5\times3}$ is the projection of the output dimensions into the state space, $\vec{0} \in [0]^P$ is the zero vector or origin of the output space, $\mathcal{D} \in \mathbb{R}^{P\times P}$ is a diagonal matrix with the maximum values of each of the output dimensions on the diagonal, $\mathbf{O} \in \mathbb{R}^{K\times P}$ is the matrix that holds the grid values of the reachable output space, $\mathbf{S} \in [-1,1]^{K\times3}$ are the dimensions of the reachable state space, $\boldsymbol{\omega}_{\text{out}} \in \mathbb{R}^{3\times P}$ is the matrix of weights of the RNN's output layer, $P = 5$ is the number of output dimensions, and K is the number of points used to map the output hypercube. We corroborated these attractor locations with the observed destinations of the trajectories; we systematically chose different initial conditions in the state space and iterated the trajectories for 100 steps. We thus determined that trajectories converge towards the attractor associated with their most dominant personality trait. If a customer's spending behavior changes such that a different personality trait becomes dominant, their trajectory changes direction towards the new appropriate attractor. Figure 3 shows these attractors in the RNN state space, with the extended trajectories converging towards their corresponding attractors.

There are three point attractors for the personality trait conscientiousness, towards which trajectories converge depending on their initial conditions. Agreeableness, extraversion, and neuroticism each have a single line attractor, while trajectories that classified as openness converge towards a single point attractor. There is no distinction in significance between attractor types of the same class, in this case fixed attractors, nor is there a significance in the fact that one personality trait corresponds to three distinct point attractors (Ceni et al., 2019). Each basin of attraction forms a cluster of trajectories, which each form a hierarchy of sub-clusters along successive levels of dominance of personality traits. This is the interpretation of the trajectory dynamics.
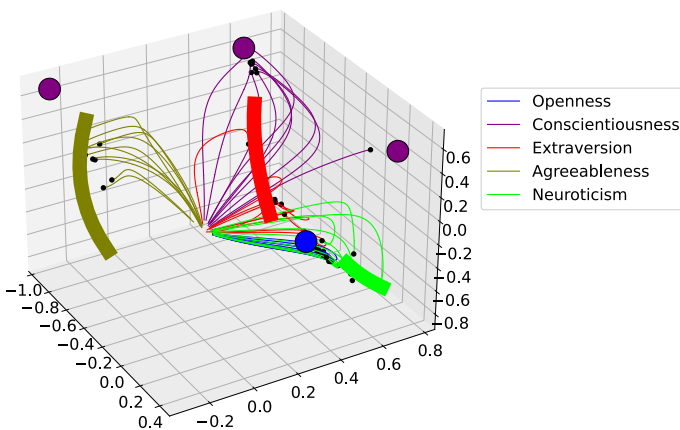

Fig. 3 The locations of a set of attractors in the state space of a RNN. There are point and line attractors that are labelled according to the customers' corresponding dominant personality traits. We show 100 trajectories, with different initial conditions, asymptotically converging to their corresponding attractors

3.2 Learning prototypical investment strategies

Maree and Omlin (2022b) showed that interpretable RL can be used for investment that matches personality. In this preliminary study, we had trained multiple RL agents to invest monthly contributions in different financial asset classes: stocks, property, savings accounts, mortgage curtailment, and luxury items. While the investment classes stocks, property, and savings accounts are self-explanatory, we define mortgage curtailment as payments that reduce the principal balance of the loan, and luxury items as items defined in, e.g., the Knight-Frank luxury investment index (Knight Frank Company, 2022). There exists a trade-off in the allocation of funds between, e.g., mortgage curtailment and purchasing stocks: there are clear differences in expected risk and reward between these two strategies, which may appeal differently to different personality types. We obtained asset prices from the S &P 500 index (Yahoo Finance, 2022), the Norwegian property index (Statistics Norway, 2022), and the Norwegian interest rate index (Norges Bank, 2022) for the period between 1 January 1992 and 31 December 2021. We indexed these prices relative to their values on 1 January 1992 and plot these indices in Fig. 4.

With the help of a panel of banking experts from a major Norwegian bank, we ranked these asset classes according to a set of desirable asset class properties: high expected long-term returns, high perceived asset liquidity, low capital prerequisite, low expected long-term risk, and high perceived novelty. We based our assessment on known characteristics of each personality trait; (1) openness that values novelty and is drawn to change; (2) conscientiousness that is predisposed to planning and values structure; (3) extraversion that values having interesting topics for discussion; (4) agreeableness that values contributing to society; and (5) neuroticism that can more easily experience stress and anxiety (Tauni et al., 2017; Rizvi & Fatima, 2015). Our experts considered the relative affinities that each personality trait might have towards each of the asset class properties; they associated the personality traits with these properties, as shown in Table 1.
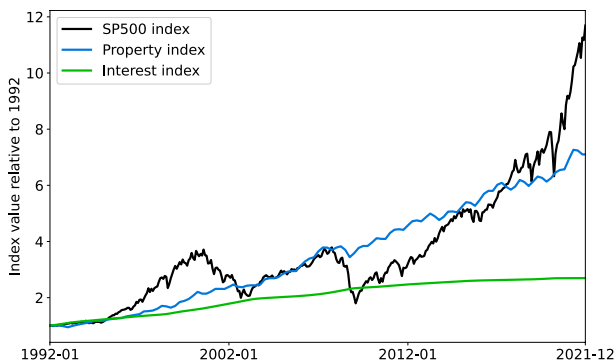

Fig. 4 Asset pricing data for the S &P500 index, Norwegian property index, and Norwegian interest rate index. The values are relative to their respective values on 1 January 1992. These values are used in the state observations of our RL agents

Table 1 Matrix *A* containing a set of asset class properties and their associations with the five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism

Asset class property	Open.	Cons.	Extra.	Agree.	Neur.
High returns	1	1	2	1	1
High liquidity	2	− 1	2	1	2
Low capital prerequisite	0	− 1	1	1	1
Low risk	− 1	2	− 1	1	2
High novelty	2	0	2	0	− 1

The values are in the set $\{n \in \mathbb{Z} \mid -2 \leq n \leq 2\}$ and indicate a strong negative, slightly negative, neutral, slightly positive and strong positive association, respectively

Table 2 Matrix *B* containing ratings for the asset classes with regard to a set of properties

Asset class property	Savings	Property	Stocks	Luxury	Mortgage
High returns	0.25	0.67	1.00	0.05	0.50
High liquidity	1.00	0.25	0.80	0.10	0.05
Low capital prerequisite	0.80	0.25	1.00	0.50	1.00
Low risk	1.00	0.32	0.10	0.05	1.00
High novelty	0.10	0.25	0.75	1.00	0.10

The values are in the range [0, 1] and higher values represent higher performance in each of the asset class properties

Table 3 Coefficients, in the range [−1, 1], associating asset classes to prototypical personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism

Asset type	Open.	Cons.	Extra.	Agree.	Neuro.
Savings account	− 0.11	0.08	− 0.15	0.51	0.68
Property funds	− 0.15	0.32	− 0.22	− 0.36	− 0.24
Stock portfolio	0.82	− 0.61	0.95	0.42	0.12
Luxury expenses	0.16	− 0.51	− 0.07	− 0.80	− 0.81
Mortgage repayments	− 0.72	0.72	− 0.52	0.23	0.25

Higher values indicate where personality traits might have higher affinities towards asset classes

The result showed that, for instance, the openness trait might highly value asset liquidity and novelty; because of their openness to new experiences, they might prefer to have cash readily at hand when such an opportunity presents itself, or they might value assets that in themselves may be perceived as novel.

Another example is that the conscientiousness trait might prefer assets with low risk. The same panel of experts then ranked the asset classes according to the same set of properties, which we show in Table 2. We quantified risk and return from historical price data and the Sharpe ratio, respectively, and the values in Table 2 are normalized from these results.

We calculated a set of coefficients *C* that associate asset classes with personality traits using matrix multiplication: $C = (A^{\mathrm{T}} \cdot B^{\mathrm{T}})^{\mathrm{T}}$. These coefficients, scaled to the

range $[-1, 1]$ and shown in Table 3, quantify personality-based affinities towards different asset classes.

These coefficients reveal that, for example, the extraversion trait has a high preference for stocks, whereas the conscientiousness agent prefers a combination of mortgage curtailment and property investment. This is in line with the findings of Gladstone et al. (2019) and Ramon et al. (2021). When scaled so that they add up to one and their minimum values are zero, these coefficients become the regularization priors π_0 in Eq. (1); we regularized the objective functions of five prototypical agents to instill intrinsic affinities for certain asset classes. Each agent learned an investment strategy associated with one of the five personality traits, which is the interpretation of their policies. Figure 5 shows these strategies, where each agent acted in an environment in which it invested a fixed monthly amount of 10,000 Norwegian Kroner (NOK) over 28 years. The data included 30 years' pricing history between 1992 and 2022, but the first 24 months were used to initialize the RL environment variables: moving average convergence divergence (MACD) and relative strength index (RSI). Investments therefore started in 1994.

These prototypical agents clearly learned unique investment strategies. For simplicity, we did not include transaction costs, since investment happened with fixed amounts and frequencies (monthly); transaction costs are negligible and equal across comparisons. The openness agent initially preferred luxury items, in line with their openness to new experiences, and later purely invested in stocks, which had scored high in novelty. In contrast, the conscientiousness agent preferred to reduce risk through property investment, followed by a resolute mortgage curtailment. The prototypical agents' affinities and their long-term strategies are independent of market conditions and the duration of the investment period, because they are defined by constant priors (see Fig. 5). These are the low-level policies that we intend to
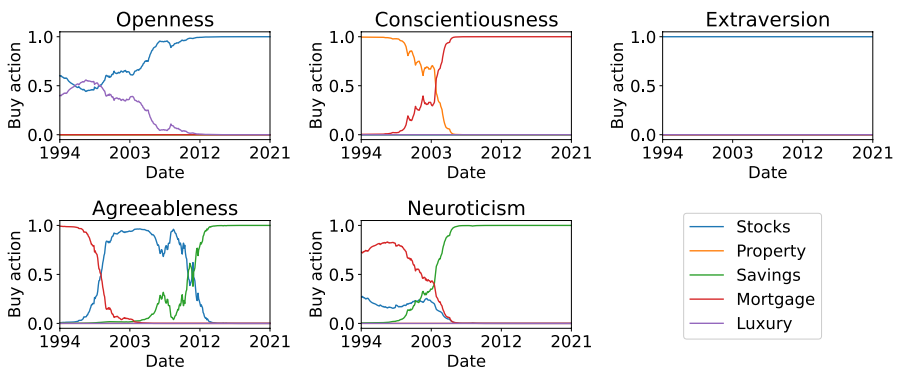

Fig. 5 Action distributions of the five prototypical agents over a 30 year time period: between the ages of 30 and 60. Each figure represents the investment actions taken by one of the prototypical agents, who each associates with a single personality trait. Each line represents the fractional monthly investment into a labelled class of assets across the time period, e.g. the conscientiousness agent initially invests solely in property and subsequently in mortgage curtailment, while the extraversion agent consistently invests the entire monthly amount in stocks. A declining trend does not indicate selling of assets but rather a reducing monthly investment amount; the values on the y-axes are strictly positive indicating our agents never sell assets but rather change their monthly investment distributions

orchestrate into personalized investment strategies; customers have varying degrees of membership in each of the five personality traits, resulting in unique preferences for different asset classes that may change over time.

3.3 Hierarchical orchestration and temporal composition

We are inspired by the premise that there is a causal relationship between personality-matched spending and happiness (Matz et al., 2016). Tauni et al. (2017) provides empirical evidence that correlates personality to stock trading behavior, confirming earlier results from Rizvi and Fatima (2015). We, therefore, extend the concept of spending behavior in time to prosperity management. Our goal is to learn, through high-level RL orchestration, the optimum composition that match customers' unique financial personalities. Our RL agent orchestrates the actions of low-level prototypical agents according to customers' extracted behavioral trajectories (Fig. 2). With actions adding up to one, representing the fraction of the investment amount allocated to each low-level agent, it maximizes the following reward function:

$$R = \vec{H} \cdot \left(\vec{P} \cdot \mathbf{C} \right), \tag{2}$$

which is the dot product between the current values of asset class holdings $\vec{H}$ and the customer's preference for each asset class. This preference is the dot product of the customer's personality vector $\vec{P}$, i.e., the set of five values representing their degrees of membership in each of the personality traits, and the set of coefficients $\mathbf{C}$ that relate each asset class with each personality trait (Table 3). The dot product is a scalar value that represents a customer's association with each asset class. By adding the associations of each personality trait with the different asset classes, multiplied by a customer's fuzzy degree of membership in the personality trait, we estimate the customer's association with each asset class. This reward measures the correlation between spending behavior and investment strategy, which we call the *satisfaction index*; the higher the satisfaction index, the higher the correlation between spending behavior and investment strategy. A limitation of this metric is that it is not a fair performance comparison of different customers with different personality profiles; the satisfaction indices will be different between one customer with a perfectly conscientious profile and portfolio and one with a perfectly extraverted profile and portfolio. It is, however, a metric that enables comparison between different methods of composing a strategy for a given customer, and that is how we use it. We then use the regularization prior:

$$\pi_0 = \frac{\vec{P}}{\sum \vec{P}}, \tag{3}$$

to instill an intrinsic RL affinity in a set of DDPG agents. The actor consisted of three vanilla RNN nodes and an output layer of five actions with a softmax activation. The critic had a similar three-node RNN layer for the states which, concatenated with the actions, were succeeded by a 1000 node feed-forward layer and a single output

node with no activation function. We found that three RNN nodes consistently provided high total rewards, which is consistent with findings that RNN architectures generally perform well in low-dimensional representations (Maheswaranathan et al., 2019). We tuned our hyperparameters using a one-at-a-time parameter sweep to reach the following optima: the actor and critic learning rates were 0.005 and 0.01, respectively, the target network update parameter τ was 0.05, the discount factor γ was 0.95, and the regularization scaling factor λ was 5.

Finally, we trained a RNN to predict this composition of prototypical agents from a sample of 500 pre-trained orchestration agents; the orchestration agents learned their strategies, as described above, from the data for 500 unique customers of a major Norwegian bank. We used the customers' feature trajectories—their encoded spending behavior—as input to a neural network with three RNN nodes. The output from this network is the actions of the orchestration agents, i.e. the unique, locally optimal combination of the five prototypical agents. We used 400 customers for training and 100 for testing. The learning rate was 0.0005 and the model typically converged within 10,000 iterations. We used this model to predict the composition of prototypical agents for customers as their spending behavior varies in time; the RNN uses a rolling window of 6 years' spending behavior.

4 Results

In this section, we compare hierarchical RL to simple linear combinations of the prototypical agents; our agents find locally optimum compositions with similar financial returns, but improved personalization compared to simple linear combinations. We also demonstrate how these compositions can accommodate changing spending behavior in time; financial personalities may fluctuate in time, and our system adapts in a non-erratic way.
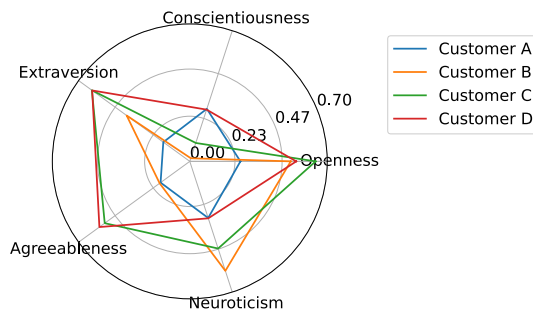

Fig. 6 The personality vectors representing the personality traits of four real customers. Each colored line represents a customer and each axis on the radar plot represents a personality trait. The values on the axes are in the range [0, 1] and represent the customers' degree of membership in each of the personality traits. These customers were selected to represent a range of personality profiles: Customer A has a balanced profile, Customer B scores high in neuroticism and openness, Customer C scores high in openness and extraversion, and Customer D scores high in extraversion, agreeableness, and openness

4.1 Hierarchical orchestration of prototypical agents

For illustration purposes, we selected four customers from a major Norwegian bank for whom we trained personal orchestration agents; their personality vectors, visualized in Fig. 6, were derived from their historical financial transactions using the RNN described in Sect. 3.1. They were chosen to represent a range of personality profiles.

Customer A has a relatively balanced profile, with low variation in the values of their personality vector, which also has relatively small values. This contrasts with Customer B who scores high in neuroticism and openness, Customer C who scores high in openness and extraversion, and Customer D who scores high in extraversion, agreeableness, and openness. Their respective regularization priors are shown in Table 4.

The regularization prior for Agent A $\pi_{0,A}$ (the agent for Customer A) consequently has a low variation in its values while $\pi_{0,B}$ assigned the highest weight to neuroticism and openness, $\pi_{0,C}$ assigned the highest weight to openness and extraversion, and $\pi_{0,D}$ assigned the highest weight to extraversion, agreeableness, and openness.

These four customers' personality profiles, and consequently the orchestration agents' actions, were constant in time. Customers' personality profiles may naturally vary in time, causing directional changes in their behavioral trajectories, which alter the orchestration agent's action distribution. We will discuss the effects of time-variant customer spending behavior on the compositions in Sect. 4.2. The investment strategies for the four customers are shown in Fig. 7.

Although these strategies might seem similar, there are significant differences: Customer A never invested more than 60% of their monthly allocation in stocks, while Customer D invested up to 90% in stocks, and Customer A was the only one to invest significantly in property. This is due to Customer A having the highest relative degree of conscientiousness, i.e., they preferred a reduced risk. In contrast, Customer D had the highest risk in their portfolio by investing the least in property and mortgage curtailment and the most in stocks, due to their low score in neuroticism which increases their appetite for risk. When comparing Customers B and C, Customer B invested more in savings accounts and less in stocks in the period between 150 and 250 months. This is due to their differences in agreeableness and neuroticism, where customer B scored higher in neuroticism and lower in agreeableness. In Fig. 5, the prototypical agents associated with neuroticism and agreeableness are

Table 4 Regularization priors used during training of the orchestration agents of four customers, named A through D

Prior	Open.	Cons.	Extra.	Agree.	Neur.
$\pi_{0,A}$	0.22	0.24	0.14	0.15	0.25
$\pi_{0,B}$	0.30	0.01	0.23	0.11	0.35
$\pi_{0,C}$	0.27	0.04	0.26	0.23	0.20
$\pi_{0,D}$	0.23	0.12	0.27	0.25	0.13

Each row represents the regularization prior $\pi_{0,i}$ for one of the orchestration agents $i \in [A, D]$. The values are in the range [0, 1] and add to one for each prior. They represent the fraction of investment amount allocated to each prototypical low-level agent: openness, conscientiousness, extraversion, agreeableness and neuroticism. A higher values indicates a higher weighting of that agent's strategy

Fig. 7 Investment advice from four personal investment agents for four different customer personalities; they are the combined actions of the prototypical agents according to the orchestration agent. Each plot shows the investment advice in time for a single customer, named "Customer A" through "Customer D" in accordance with the labels in Fig. 6. Declining trends do not indicate selling of assets but rather reduced monthly investment in that asset; the values on the *y*-axes are strictly positive indicating that assets are never sold, but investment distributions change across assets

the only two to invest in savings, and the neuroticism agent started investing in savings much earlier and with higher percentages. Despite the nuanced differences in investment approaches, the general advice for all customers was similar: first pay down mortgages to reduce debt repayments, then accept higher risk with higher returns from stocks and benefit from compound growth, and finally toward retirement age reduce risk through savings accounts. This is consistent with conventional financial advice: younger people with more disposable income may accept more risk for higher returns. Very interestingly, this was not explicit in the objective function, which had no elements of risk, while the effect of compound growth was evident only in increased final returns.

The monthly investments accumulated to 3.36 million NOK, and the portfolios were initiated with a 2 million NOK property investment with a corresponding 2 million NOK mortgage; individual strategies may vary between, e.g., quickly reducing the principal balance of the loan thus avoiding interest or investing in more risky asset classes such as stocks. The theoretical maximum return was 27.7 million NOK, achieved when investing purely in stocks. The final financial returns for our four customers were very similar: after 28 years of investing 10,000 NOK per month, they all had portfolio values ranging between 21 and 24 million NOK. However, our aim was to optimize customer satisfaction in their portfolio while still achieving high returns. We note that the satisfaction index is not a suitable metric for comparing different customers, and this is reflected in the results, where satisfaction indices between customers had greater variation than their financial returns. However, we compare the satisfaction index between different compositions of prototypical agents for the same customer: Table 5 shows the results of the orchestration agents and those of a linear combination of the prototypical agents.

Table 5 Performance metrics comparing the orchestration agent to a simple linear combination of the prototypical agents

Customer	Orchestration agent			Linear combination		
	Portfolio value mill. NOK	Satisfaction index	Sharpe ratio	Portfolio value mill. NOK	Satisfaction index	Sharpe ratio
A	20.9	3.1	0.4393	20.9	3.0	0.4367
B	22.0	12.9	0.3704	21.7	12.7	0.3730
C	22.7	19.0	0.3528	22.4	17.8	0.3616
D	23.8	19.5	0.3350	22.6	14.5	0.3706

We list the resulting portfolio values and satisfaction scores for both these scenarios after investing 10,000 NOK per month for 28 years according to the strategies shown in Fig. 7. Here, the Sharpe ratio is the mean of the monthly returns divided by the standard deviation of the monthly returns

This linear combination is the dot product of the personality vector and the action vectors of the prototypical agents, scaled such that the resulting actions add up to one; the actions of the prototypical agents were weighted according to customers' personality vectors. In terms of profit and satisfaction index, the orchestration agent never performs worse than a linear combination of prototypical agents; although it typically achieves only slightly better financial returns, it can significantly improve the satisfaction index. This was not the case when using feed-forward networks to process the customer spending input, which returned inconsistent results across multiple training runs and frequently performed worse than the simple linear combination. This is consistent with findings from (Tovanich et al., 2021) that spending patterns in time hold salient information not evident in non-temporal data. The Sharpe ratios are similar between the customers' orchestration agents and linear combinations, and only Customer A had a higher Sharpe ratio for the orchestration agent. This is explained by Customer A's relatively high score in conscientiousness which, as stated before, resulted in increased investment in property—a lower risk asset class—and a corresponding reduction in portfolio risk. We calculated the Sharpe ratio for the global optimum strategy—investing solely in stocks—as 0.2856 indicating an increased risk in the portfolio. This strengthens our argument that our locally optimal personalized strategies could be improvements over the global optimum in returns.

We regularized the orchestration agents to act according to a specified prior with the same action distribution as for the linear combination scenario. Through stochastic gradient descent, they optimized the satisfaction index in that region of the action space. In Fig. 11, we illustrate the policy convergence towards local optima of each of the four orchestration agents. The policies were randomly initialized, but quickly converged to local optima in close proximity to the regularization priors in the action space. The learned strategies are thus interpretable.

4.2 Time-variant analysis

We have access to historical transactions dating back a maximum of 6 years, which hinders long-term time-varying analyses of customers' spending behavior. However,

we created a fictitious customer, Customer E, by copying the financial transactions of two distinct customers: one who scored high in conscientiousness and another who had a slightly more balanced profile while demonstrating mostly extraverted spending behavior. We constructed Customer E's transaction history as follows: we duplicated 1 year's transactions from the conscientious customer 10 times, then 1 year's transactions from the extraverted customer 10 times, and the final 8 years' transactions again from the conscientious customer. Customer E thus exhibited 10 years of conscientious spending behavior, followed by 10 years of mixed, but mostly extraverted behavior, followed by the final 8 years of contentious behavior once again. Our aim was to demonstrate what effect a change in spending behavior has on the investment strategy. Figure 8 shows the encoded spending behavior, or the feature trajectory, of this fictitious customer. It follows the expected behavior and converges towards the corresponding conscientiousness and extraversion attractors. It is not expected that a trajectory converges exactly on top of an attractor with every change in spending behavior but that it moves towards the corresponding attractor. This illustrates the interpretation of our feature extraction model: by observation and with knowledge of the locations of the attractors that govern the dynamics of the system, we can reason about the functioning of the model.

We trained a RNN from the spending behavior of 500 customers from a major Norwegian bank to predict the actions of their corresponding orchestration agents. Using this RNN, we predicted the recommended composition of prototypical agents for Customer E, shown in Fig. 9. This investment strategy highly favors the conscientiousness agent in the first 10 years, after which the composition changes to a mixture of agents that is biased towards extraversion. This transition does not happen immediately and there is a gradual shift over the course of a few years. This is important, as financial advice should not be erratic. The mixture of agents is
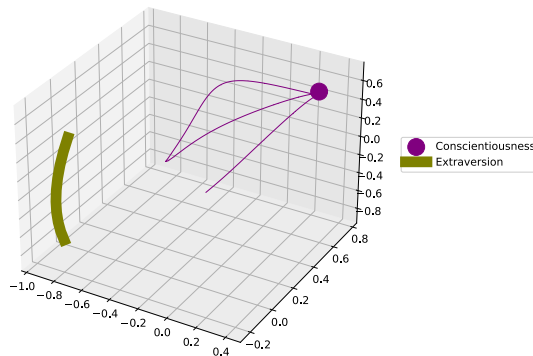

Fig. 8 The encoded spending behavior for a fictitious customer, Customer E, drawn as feature trajectories in the state space of the RNN from Fig. 2. This customer's financial transactions were such that their spending personalities were first predominantly conscientious, then extraverted, and finally conscientious once again. We show the two corresponding attractors and the customer's trajectory which initially converges on the conscientiousness attractor. As soon as the customer's spending pattern changes, the trajectory moves towards the corresponding new attractor: extraversion. Finally, and before a sufficient time has passed for the trajectory to converge on the new attractor, the spending pattern changes back to conscientiousness and the trajectory once again converges on that attractor
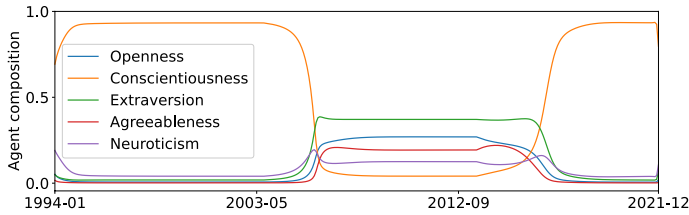
Fig. 9 The recommended composition of prototypical agents for Customer E. We created Customer E to display highly conscientious spending behavior between 1994 and 2004. Between 2005 and 2015, they displayed spending behavior related to a mixed personality profile which was mostly extraverted. From 2015 onward, their spending was once again conscientious. This time-varying spending behavior is reflected in the weights assigned in the composition of prototypical agents: conscientious spending behavior results in a conscientious investment strategy, which can change in time with changing spending behavior

expected and can be explained by observing the spending trajectory in Fig. 8: the trajectory has not yet converged to the extraversion attractor and may fall close to the basin of attraction of several other personality attractors. It also corresponds to the behavior of the selected customer from whom we copied transactions: they were predominantly extraverted but also showed behavior from other traits such as openness, agreeableness, etc. This result shows that while the dominant personalty trait is important—extraversion is the largest portion of the composition—our system also considers other traits. In the last 8 years, the composition shifts back to favoring the conscientiousness agent.

Figure 10 shows the composed strategy for Customer E which, unsurprisingly, closely follows the prototypical conscientiousness strategy in the initial and final phases, while in the middle it invests more in stocks and savings accounts. We show the portfolio value and asset class holdings in Figs. 12 and 13 respectively. While
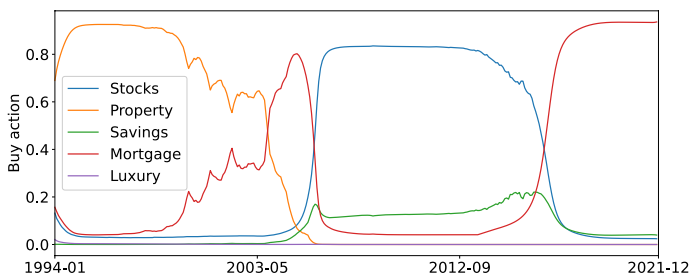

Fig. 10 The long-term, time-variant investment strategy for a fictitious customer, Customer E. We created Customer E to display highly conscientious spending behavior between 1994 and 2004. Between 2005 and 2015 they displayed spending behavior related to a mixed personality profile which was mostly extraverted. From 2015 onward their spending was once again conscientious. The investment strategy, according to the time-variant composition of the prototypical agents (Fig. 9), is related to the customer's spending behavior in time. Initially, the conscientious spender invests conscientiously—in low risk asset classes, namely property—while between 2005 and 2015 the extraverted spender invests mostly in stocks with an element of agreeableness evident in their investment in savings accounts. Finally, the strategy reverts to a conscientious behavior and resolute mortgage curtailment

investment in stocks corresponds to strategies from extraversion, openness, and agreeableness, investment in savings are related to the agreeableness strategy. This strategy is clearly interpretable from the perspective of spending behavior in time. From a customer's financial records, we can estimate their spending personality and extract behavioral features using an RNN. We can reason about these features based on their locations and trajectories in the state space of our RNN.

Then, we can combine the actions of five, interpretable, prototypical agents to suggest an investment strategy. We can also reason about this strategy given the inherent affinities of our prototypical agents. This ability to reason about the predictions of a system inspires trust and removes a cloak of uncertainty.

5 Conclusions

Machine learning is essential for personalizing financial services. Its acceptance is contingent on understanding the underlying models, which makes model explainability and interpretability imperative. Our reinforcement learning model blends investment advice that is aligned with different personality traits. Its interpretation follows from the global intrinsic affinities of the learned policies, i.e., affinities that are independent of the current state. These policies not only result in good profit, but also similar profits are achieved across different personality profiles despite their distinct strategies. For instance, they avoid risk for highly conscientious individuals, while pursuing novelty for individuals that are more open to new experiences. Their time-variant strategies adapt in a non-sporadic way to changes in spending behavior. Interestingly, our agents have learned the concept of risk without this being explicit in the objective function. Across all portfolios, the advice is consistent with conventional wisdom: younger investors may accept higher risk, which typically reduces with age. It remains to be seen whether this is simply a consequence of optimizing profit while balancing the intrinsic action distribution, or whether our agents have learned deeper strategies of asset management. In future work, we intend to investigate this phenomenon by extracting an explanation for our agents' decisions. It will also be interesting to extend our method to local intrinsic affinity, where the preferred policy also depends on the *current state*. It is compelling to generalize the approach by Nangue Tasse et al. (2020) who decompose tasks and suggest a Boolean algebra for the composition of the learned strategies; ours is a fuzzy composition of prototypical agents that might benefit from such an extension. The potential applications for our method go beyond investment advice and include, e.g., autonomous vehicles, personalized teaching and learning, treatment of chronic diseases, or the design of virtuous agents in the context of artificial morality.

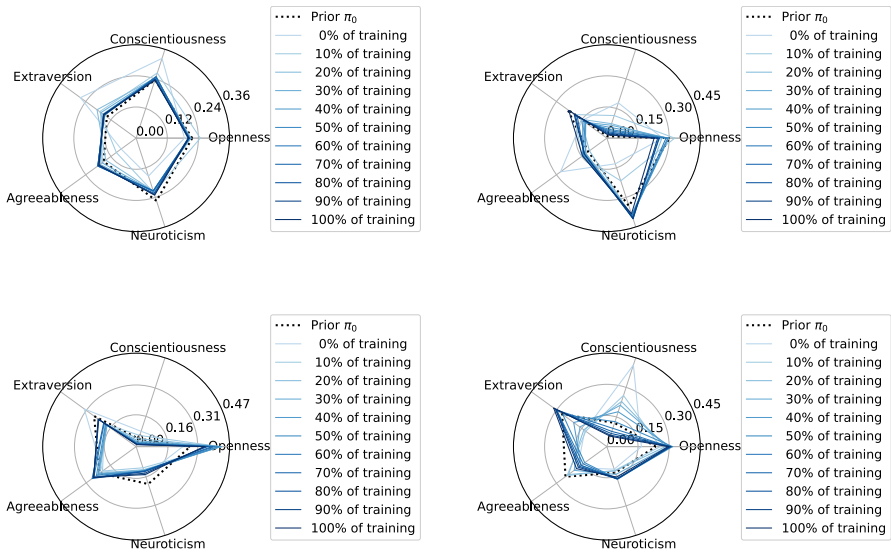## Appendix 1: Training convergence

See Fig. 11.

Fig. 11 Training convergence of the orchestration agents for four different customer personality profiles. Each successively darker blue line represents the orchestration action distribution after an increasing number of training runs. As training progresses, the successively darker blue lines converge towards the learned action distribution. The black dotted line represents the regularization prior $\pi_{0,i}$. The figures show how randomly initialized policies converge towards their specified priors and settle in a local optima in close proximity

Appendix 2: Example portfolio returns

See Figs. 12 and 13.
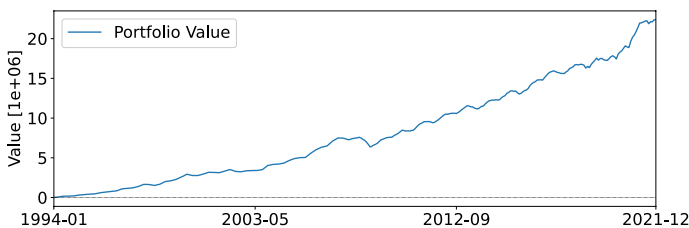
Fig. 12 The resulting portfolio value for Customer E, a fictitious customer designed to illustrate the time-varying investment strategy for a customer whose spending behavior varies in time. Customer E first exhibited conscientious spending behaviour, followed a period of extraverted behavior with significant elements from other traits, and finally they reverted to conscientious spending. The portfolio value follows an upward trend with a slight downward variability in about 2008. The reason for this contraction becomes evident when combining information from Figs. 13 and 4: the customer has a relatively high holding in property for which there was a market contraction in 2008
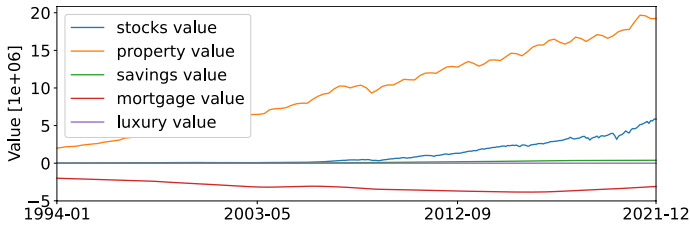
Fig. 13 The resulting portfolio value for Customer E, a fictitious customer designed to illustrate the time-varying investment strategy for a customer whose spending behavior varies in time. Customer E first exhibited conscientious spending behaviour, followed a period of extraverted behavior with significant elements from other traits, and finally they reverted to conscientious spending. The asset class holdings correspondingly favours property initially and this asset class experiences compound growth throughout the investment period following its index shown in Fig. 4. The strategy only invests in stocks between about 2005 and 2017 (refer to Fig. 10) and the stock holding is correspondingly low

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval Not applicable.

Consent to participate Personal data were anonymized and processing was done on the basis of consent in compliance with the European General Data Protection Regulation (GDPR).

References

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

Apeh, E. T., Gabrys, B., & Schierz, A. (2011). Customer profile classification using transactional data. *2011 Third World Congress on Nature and Biologically Inspired Computing* (pp. 37–43). Salamanca, Spain.

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

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115.

Beyret, B., Shafti, A., & Faisal, A. (2019). Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 5014–5019). Macau, China.

Cao, L. (2021). AI in finance: Challenges, techniques and opportunities. *Banking & Insurance eJournal, 55,* 1–38.

Ceni, A., Ashwin, P., & Livi, L. F. (2019). Interpreting recurrent neural networks behaviour via excitable network attractors. *Cognitive Computation, 12,* 330–356.

Fernández, A. (2019). Artificial intelligence in financial services. Tech. rep., The Bank of Spain, Madrid, Spain.

Galashov, A., Jayakumar, S., Hasenclever, L., et al. (2019). Information asymmetry in KL-regularized RL. *International Conference on Learning Representations (ICLR)* (pp. 1–25). New Orleans: Louisiana, United States.

García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research, 16*(42), 1437–1480.

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science, 30*(7), 1087–1096.

Hengst, B. (2010). *Hierarchical reinforcement learning* (pp. 495–502). Springer.

Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems, 214*(106685), 1–24.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Knight Frank Company (2022) Knight Frank luxury investment index. https://www.knightfrank.com/wealthreport/luxury-investment-trends-predictions/. Accessed 27 May 2022.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., et al. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 1–9). Curran Associates Inc.

Levy, A., Platt, R., & Saenko, K. (2019). Hierarchical reinforcement learning with hindsight. In: *International conference on learning representations* (pp. 1–16).

Lillicrap, TP., Hunt, JJ., & Pritzel A., et al. (2019). Continuous control with deep reinforcement learning. arXiv:1509.02971.

Maheswaranathan, N., Williams, A. H., Golub, M. D., et al. (2019). Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Advances in Neural Information Processing Systems (NIPS), 32,* 15696–15705.

Maree, C., & Omlin, C. W. (2021). Clustering in recurrent neural networks for micro-segmentation using spending personality. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–5)

Maree, C., & Omlin, C. W. (2022). Understanding spending behavior: Recurrent neural network explanation and interpretation (in print). In: *IEEE computational intelligence for financial engineering and economics* (pp. 1–7)

Maree, C., & Omlin, C. (2022). Reinforcement learning your way: Agent characterization through policy regularization. *AI, 3*(2), 250–259.

Maree, C., & Omlin, C. W. (2022). Can interpretable reinforcement learning manage prosperity your way? *AI, 3*(2), 526–537.

Marzari, L., Pore, A., Dall'Alba, D., et al. (2021). Towards hierarchical task decomposition using deep reinforcement learning for pick and place subtasks. *20th international conference on advanced robotics (ICAR)* (pp. 640–645). Ljubljana, Slovenia.

Matz, S. C., Gladstone, J. J., & Stillwell, D. (2016). Money buys happiness when spending fits our personality. *Psychological Science, 27*(5), 715–725.

Millea, A. (2021). Deep reinforcement learning for trading: A critical survey. *Data, 6*(11), 1–25.

Milnor, J. (2004). *On the concept of attractor* (pp. 243–264). Springer.

Miryoosefi, S., Brantley, K., & Daume, III H., et al. (2019). Reinforcement learning with convex constraints. In: *Advances in neural information processing systems* (pp. 1–10).

Mousaeirad, S. (2020). Intelligent vector-based customer segmentation in the banking industry. arXiv: 2012.11876.

Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. *34th conference on neural information processing systems (NeurIPS 2020)* (pp. 1–11). Vancouver, Canada.

Norges Bank. (2022). Interest rates. https://app.norges-bank.no/query/#/en/interest. Accessed 30 Jan 2022.

Pateria, S., Subagdja, B., Tan, Ah., et al. (2021). Hierarchical reinforcement learning: A comprehensive survey. *Association for Computing Machinery, 54*(5), 1–35.

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

Rizvi, S., & Fatima, A. (2015). Behavioral finance: A study of correlation between personality traits with the investment patterns in the stock market. *Managing in Recovering Markets* (pp. 143–155). New Delhi: Springer India.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing, 21*(1), 3–8.

Statistics Norway. (2022). Table 07221-Price index for existing dwellings. https://www.ssb.no/en/statbank/table/07221/. Accessed 30 Jan 2022.

Stefanel, M., & Goyal, U. (2019). *Artificial intelligence & financial services: Cutting through the noise*. APIS partners, London, England: Tech. rep.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.

Tauni, M. Z., Rao, Zu. R., Fang, H., et al. (2017). Do investor's big five personality traits influence the association between information acquisition and stock trading behavior? *China Finance Review International, 7*(4), 450–477.

Tovanich, N., Centellegher, S., Bennacer Seghouani, N., et al. (2021). Inferring psychological traits from spending categories and dynamic consumption patterns. *EPJ Data Science, 10*(24), 1–23.

Vieillard, N., Kozuno, T., & Scherrer, B., et al. (2020). Leverage the average: An analysis of KL regularization in reinforcement learning. In: *Advances in Neural Information Processing Systems (NIPS)* (vol. 33, pp. 12163–12174). Curran Associates.

Yahoo Finance. (2022). Historical data for S &P500 stock index. https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC. Accessed 30 Jan 2022.