NBER WORKING PAPER SERIES

MODEL-FREE AND MODEL-BASED LEARNING AS JOINT DRIVERS OF INVESTOR BEHAVIOR

Nicholas C. Barberis
Lawrence J. Jin

Working Paper 31081
http://www.nber.org/papers/w31081

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2023

We are grateful to Andrew Caplin, Alex Chinco, Cary Frydman, Chen Lian, Elise Payzan-LeNestour, Antonio Rangel, Josh Schwartzstein, Andrei Shleifer, Michael Woodford, and seminar participants at Arizona State University, Baruch College, Caltech, Columbia University, Cornell University, Harvard University, Imperial College London, Texas A&M University, Tsinghua University, the University of California Los Angeles, the University of Pennsylvania, Washington University, Yale University, the AFA Annual Meeting, the Behavioral Economics Annual Meeting, the Miami Behavioral Finance conference, the NBER Behavioral Finance and Behavioral Macroeconomics conferences, and the Sloan-Nomis School on Cognitive Foundations of Economic Behavior for very useful feedback. We are also grateful to Colin Camerer, Nathaniel Daw, Peter Dayan, Sam Gershman, John O'Doherty, and members of their lab groups for very helpful discussions about the psychological concepts in the paper. Steven Ma provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Model-free and Model-based Learning as Joint Drivers of Investor Behavior
Nicholas C. Barberis and Lawrence J. Jin
NBER Working Paper No. 31081
March 2023
JEL No. D03,G02,G11

ABSTRACT

In the past decade, researchers in psychology and neuroscience studying human decision-making have increasingly adopted a framework that combines two systems, namely "model-free" and "model-based" learning. We import this framework into a simple financial setting, study its properties, and use it to account for a range of facts: facts about investor behavior, such as extrapolative demand and experience effects; facts about beliefs, such as overreaction in beliefs and the relationship between beliefs and stock market allocations; and facts about asset prices, such as excess volatility. More broadly, the framework offers a way of thinking about individual behavior that is grounded in recent evidence on the computations that the brain undertakes when estimating the value of a course of action.

Nicholas C. Barberis
Yale School of Management
P O Box 208200
New Haven, CT 06520-8200
and NBER
nick.barberis@yale.edu

Lawrence J. Jin
Johnson College of Business
Cornell University
Ithaca, NY 14850
United States
lawrence.jin@cornell.edu

A data appendix is available at http://www.nber.org/data-appendix/w31081

1 Introduction

A fundamental question in both economics and psychology asks: How do people make decisions in dynamic settings? The traditional answer in economics is to say that people act as if they have solved a dynamic programming problem. By contrast, over the past decade, psychologists and neuroscientists have embraced a different framework for thinking about decision-making in dynamic settings. This framework combines two algorithms, or systems: a "model-free" learning system and a "model-based" learning system. In this paper, we import this framework into a simple economic setting – a portfolio-choice problem where investors allocate between a risk-free asset and a risky asset – study its properties, and show that it is helpful for thinking about a range of facts in finance.[1]

The goal of both the model-free and the model-based algorithms is to estimate the value of a given action. The model-free system goes about this in a way that is different from traditional economic models in that, as its name suggests, it does not use a "model of the world": it makes no attempt to construct a probability distribution of future outcomes. Rather, it learns by experience. At each date, it tries an action, observes the outcome, and then updates its estimate of the value of the action by way of two important quantities: a reward prediction error – the reward it observes after taking the action relative to the reward it anticipated – and a learning rate. If the prediction error is positive, the algorithm raises its estimate of the value of the action and is more likely to repeat the action in the future; if the prediction error is negative, it lowers the estimated value of the action and is less likely to try it again. This model-free framework has been increasingly adopted by psychologists and neuroscientists because of evidence that it reflects actual computations performed by the brain: numerous studies have found that neurons in the brain encode the reward prediction error used by model-free learning.[2]

The model-based algorithm, by contrast, is similar to traditional economic approaches in that it does construct a model of the world – a probability distribution of future outcomes – and then uses this to compute the value of different actions. There are a number of model-based approaches; we use one that is often adopted in research in psychology and that, like the model-free system, has neuroscientific support. Under this approach, after observing an outcome at some moment in time, the model-based system increases the probability it

[1] An early paper on this framework is Daw, Niv, and Dayan (2005). Two prominent implementations in laboratory settings are Glascher et al. (2010) and Daw et al. (2011). Useful reviews include Balleine, Daw, and O'Doherty (2009) and Daw (2014). We discuss the behavioral and neural evidence for the framework in more detail in Section 2.

[2] See Montague, Dayan, and Sejnowski (1996), Schultz, Dayan, and Montague (1997), McClure, Berns, and Montague (2003), and O'Doherty et al. (2003), among many others.

assigns to that outcome and downweights the probabilities of other outcomes. To do the updating, it again uses a learning rate and a prediction error that measures how surprising a realized outcome is; as before, there is evidence that the brain computes such prediction errors (Glascher et al., 2010).

Recent research in psychology argues that, to make decisions, people use these two systems in combination: they take a weighted average of the model-free and model-based estimates of the value of different actions and use the resulting "hybrid" estimates to make a choice (Glascher et al., 2010; Daw et al., 2011).

In this paper, we import this framework into an economic setting, study its properties, and use it to account for a range of empirical facts. We pay particular attention to the model-free system – for economists, the more novel part of the framework – and on how its predictions differ from those of the model-based system. While the framework can be applied in many economic domains, it is natural to consider an application in finance: these algorithms are designed to select actions that maximize future rewards, and financial markets are an important source of reward. Within the realm of finance, we choose a simple setting: a portfolio-choice problem where an individual allocates money between a risk-free asset and a risky asset in order to maximize the expected log utility of wealth at some future horizon. This problem fits the canonical context in which model-free and model-based algorithms are applied.

We begin by analyzing the properties of our framework. Specifically, we look at how the stock market allocation proposed by each of the model-free and model-based systems depends on past stock market returns. The model-based allocation puts weights on past market returns that are positive and that decline for more distant past returns. We find that the model-free system also recommends an allocation that puts positive weight on past returns, and show that it does so through a mechanism that is new to financial economics. In brief: A good stock market return reinforces the investor's previous allocation – it raises the model-free estimate of the value of the allocation and thus encourages the investor to continue with this allocation – whether the allocation was low or high. However, this reinforcement is stronger when the prior allocation is high: for a given market return, the reward, or portfolio return, is higher when the prior allocation is high. As a consequence, on average, a good stock market return leads the investor to subsequently take a higher allocation.

We also find that, relative to the model-based system, whose recommended allocation puts heavy weight on recent returns, the model-free allocation puts substantially more weight on distant past returns. This is because it updates slowly: since it learns from experience, at each time, it updates only the value of the most recently-chosen allocation; the values of

the other allocations are unchanged and hence depend only on more distant past returns. It therefore takes a long time for the influence of past returns to fade.

We then use our framework to shed light on a range of important facts related to the aggregate stock market – facts about investor behavior, such as extrapolative demand and experience effects; facts about beliefs about future market returns and their relationship to allocations; and facts about asset prices, such as excess volatility and return predictability.

A prominent idea, motivated by empirical evidence, is that investors have extrapolative demand: their demand for a risky asset depends on a weighted average of the asset's past returns, where the weights are positive and larger for more recent returns. The analysis summarized above shows that model-free and model-based learning can both offer a foundation for extrapolative demand; the model-free system, in particular, does so in a way that is new to financial economics. Moreover, in an asset pricing setting, this extrapolative demand generates excess volatility in market returns as well as predictability in these returns. Through the model-based system, our framework preserves a role for beliefs as a driver of the excess volatility. However, through the model-free system, the framework introduces a new way of thinking about this volatility, one based on reinforcement of past actions.

Our framework also provides a foundation for experience effects – specifically, for the finding of Malmendier and Nagel (2011) that an individual's allocation to the stock market can be explained in part by a weighted average of the market returns he has personally experienced, with much less weight on returns he has not experienced. Our framework captures this because of a fundamental feature of the model-free system, namely that, because this system learns from experience, it engages only when an individual is actively experiencing rewards. As such, it puts no weight on returns an investor has not experienced. We relate our approach to thinking about experience effects to alternative, memory-based approaches (Bordalo et al., 2020; Wachter and Kahana, 2022).

Individual investors overreact to recent market returns when forming beliefs about future market returns: as shown by Greenwood and Shleifer (2014) among others, their beliefs depend strongly on recent returns even though there is little autocorrelation in realized returns. Through the model-based system, our framework captures this overreaction.

Beyond simply capturing overreaction in beliefs, our framework can also resolve two puzzling disconnects between investors' beliefs and stock market allocations. While individual investor beliefs about future stock market returns depend primarily on recent past market returns, Malmendier and Nagel (2011) find that investors' allocations to the stock market depend significantly even on distant past market returns. We reconcile these findings

by way of a deep property of our framework, which is that, of the two systems, only the model-based system has a role for beliefs: only this system explicitly constructs a probability distribution of future outcomes. When an individual is surveyed about his beliefs regarding future returns, he necessarily consults the model-based system – only this system can answer the survey question – and therefore gives an answer that depends primarily on recent past returns. However, his allocation is influenced by both the model-based and model-free systems and therefore depends significantly even on distant past returns. Through a similar mechanism, our framework can also explain another disconnect between actions and beliefs, namely the low sensitivity of allocations to beliefs documented by Giglio et al. (2021) in the cross-section of investors.

The framework can also help to account for some other empirical facts, including the large cross-sectional dispersion in investor allocations to the stock market; the individual-level inertia in these allocations over time; the widespread non-participation in the stock market among U.S. households; and the fact that many households persist in making suboptimal choices for long periods of time. We also draw out a number of predictions of the framework – for example, that for an investor who is more confident in his beliefs, the brain is likely to assign more control to the model-based system, leading the investor's allocation to be more closely tied to his beliefs.

Since the model-free system learns slowly, it is not an efficient way of making investment decisions in real time. Nonetheless, for at least two reasons, it is likely, as our paper suggests, to influence financial decision-making. First, the model-free system is a fundamental component of human decision-making. As such, it is likely to play a role in any decision unless explicitly "switched off" – and because it operates below the level of conscious awareness, many investors will not recognize its influence and will therefore fail to turn it off. Second, many people do not have a good "model" of financial markets – for example, they have a poor sense of the structure of asset returns. As a consequence, the brain is likely to assign at least some control of financial decision-making to the model-free system – again, without a person's conscious awareness – precisely because this system does not need a model of the environment.

Model-free learning algorithms are of interest not only to psychologists and neuroscientists, but also to computer scientists, albeit for a different purpose. Computer scientists see these algorithms as a powerful tool for solving challenging dynamic problems (Sutton and Barto, 2019). For example, these algorithms have been embedded in computer programs that have achieved world-beating performance in complex games such as Backgammon and Go. Psychologists and neuroscientists, by contrast, are interested in these algorithms be-

cause they see them as good models of how animals and humans actually behave. In this paper, we take the psychologists' perspective: we are proposing that these algorithms can shed light on the behavior of real-world investors.

The full name of model-free learning is model-free reinforcement learning. Reinforcement learning is a fundamental concept in both psychology and neuroscience – and, as described above, in some areas of computer science. However, it has a much smaller footprint in economics and finance, where model-based frameworks dominate instead. A central theme of this paper is that model-free learning may be more relevant in economic settings than previously realized. Nonetheless, our approach does have antecedents in economics – most notably in research in behavioral game theory on how people learn what actions to take in strategic settings (Erev and Roth, 1998; Camerer, 2003, Ch. 6). One important idea in this line of research, Camerer and Ho's (1999) experience-weighted attraction learning, combines reinforcement and model-based learning in a way that is reminiscent of the hybrid model we consider below.

Our paper is also part of a new wave of research in behavioral economics that seeks to move beyond the high-level psychological phenomena made famous by Daniel Kahneman and Amos Tversky and to instead incorporate deeper, lower-level psychological processes into economic models. This research has studied topics such as memory, attention, and perceptual coding. In this paper, our focus is on learning algorithms.[3]

In Section 2, we formalize the model-free and model-based learning algorithms and show how they can be applied in a simple economic setting. In Section 3, we present an example to show how the two algorithms work and then analyze the properties of the framework. In Section 4, we use the framework to account for a range of facts about investor behavior. Section 5 considers some extensions while Section 6 concludes.

2 Model-free and Model-based Algorithms

Researchers in the fields of psychology and neuroscience are increasingly adopting a framework that combines model-free and model-based learning (Daw, Niv, and Dayan, 2005; Daw, 2014). In this section, we describe this framework and propose a way of applying it in an economic setting. We begin by summarizing some of the evidence that motivates the framework.

[3]Examples of papers in this new wave of research are Bordalo, Gennaioli, and Shleifer (2020), Khaw, Li, and Woodford (2021), Frydman and Jin (2022), and Wachter and Kahana (2022).

2.1 Psychological background

Under the model-free system, an individual is drawn to actions that have been rewarded in the past. Under the model-based system, actions are instead derived from a model of the environment. Both systems have deep roots in psychology – the model-free system in Thorndike's (1933) "law of effect," and the model-based system in Tolman's (1948) notion of a "cognitive map," an internal representation of the environment. An emerging view in psychology is that humans use both of these systems, in combination. This view is based both on behavioral data – data on how people behave – and on neural data.

To illustrate the two types of evidence, we summarize an experiment conducted by Daw et al. (2011). In the first stage – see Figure 1 – a participant is given a choice between two options, A and B. If he chooses A, then, with probability 0.7, he is given a choice between options C and D, and with probability 0.3, a choice between options E and F. Conversely, if he chooses B in the first stage, then, with probability 0.7, he is given a choice between E and F, and with probability 0.3, a choice between C and D. After choosing between C and D or between E and F, the participant receives the reward associated with the chosen second-stage option. He repeats this task multiple times with the goal of maximizing the sum of his rewards.[4]

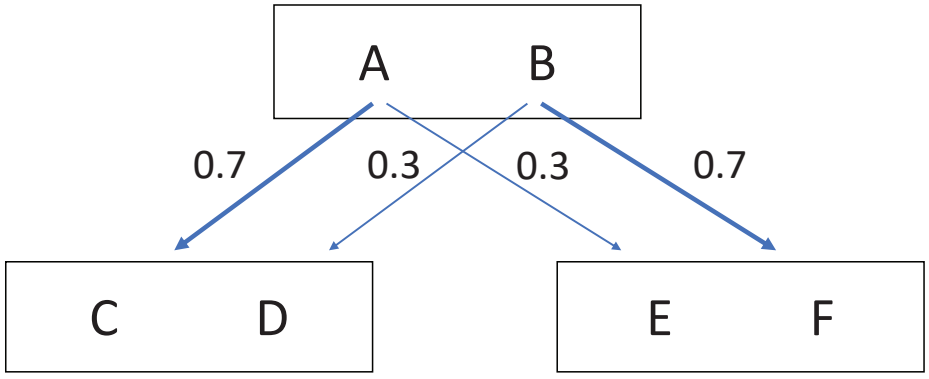

Figure 1. The diagram shows the structure of an experiment in Daw et al. (2011). In the first stage, the participant has a choice between two options, A and B; in the second stage, he chooses either between options C and D or between options E and F. The arrows indicate the transition probabilities from the first to the second stage. After making a choice at the second stage, the participant receives the reward associated with the chosen option.

[4]In the standard version of this experiment, participants are informed that each of the first-stage options is primarily associated with one of the C-D and E-F pairs but are not told which one, nor are they told the precise transition probabilities.

The model-free and model-based systems make different predictions about behavior in this setting. Suppose that the individual chooses A in the first stage and is then offered a choice between E and F; suppose that he chooses E and then receives a reward. Under the model-free system, he will be inclined to choose A again in the next trial because this choice was ultimately rewarded. Under the model-based system, however, he will be inclined to choose B in the next trial: the model-based system makes use of information about the structure of the task; since B offers a greater likelihood of ending up with the rewarded option E, he prefers B.

To evaluate the relative influence of model-free and model-based thinking on people's choices, Daw et al. (2011) run a regression of whether a participant repeats his previous first-stage choice on two variables: an indicator variable that equals one if this previous choice resulted in a reward; and this indicator interacted with another indicator variable that equals one if the individual saw the common rather than the rare second-stage options. For example, following an initial choice of A, the common second-stage options are C and D while the rare ones are E and F. If behavior is driven purely by the model-free system, only the coefficient on the first regressor will be significant. If behavior is driven purely by the model-based system, only the coefficient on the second regressor will be significant. The authors find that both coefficients are significant, which means that both systems are playing a role; an estimation exercise indicates that participants are putting approximately 60% weight on the model-free system and 40% on the model-based system.[5]

The presence of both model-free and model-based influences on behavior is also supported by neural data. The model-free and model-based systems update the values they assign to different actions using prediction errors. In an experiment similar to that of Daw et al. (2011), Glascher et al. (2010) use functional magnetic resonance imaging (fMRI) to show that neural activity in a brain region known as the ventral striatum correlates with the prediction error for the model-free system, while neural activity in an area of the prefrontal cortex correlates with the prediction error for the model-based system. These findings suggest that the brain implements the model-free and model-based algorithms when making decisions. Similar neural evidence has been reported in several other studies.[6]

We now present the formal algorithms that have been developed to capture model-free and model-based learning. In Section 2.2, we describe the model-free algorithm; in Section

[5]Charness and Levin (2005) present a different experiment in which model-free and model-based learning – in their terminology, reinforcement learning and Bayesian learning – again make different predictions. They, too, find that participant behavior is guided to a significant extent by the model-free system. More recent experimental studies with a similar theme are Payzan-LeNestour and Bossaerts (2015) and Allos-Ferrer and Garagnani (2022).

[6]We provide references to these studies later in Section 2.

2.4, we lay out a model-based learning algorithm; and in Section 2.5, we show how the two algorithms are combined. In Section 2.3, we present the portfolio-choice problem that we apply the algorithms to. For much of the paper, we will explore the properties and applications of model-free and model-based learning in this financial setting.

2.2 Model-free learning

Model-free and model-based learning algorithms are intended to solve problems of the following form. Time is discrete and indexed by $t = 0, 1, 2, 3, \ldots$ At time t, the state of the world is denoted by s_t and an individual takes an action a_t. As a consequence of taking this action in this state, the individual receives a reward r_{t+1} at time $t+1$ and arrives in state s_{t+1} at that time. The joint probability of s_{t+1} and r_{t+1} conditional on s_t and a_t is $p(s_{t+1}, r_{t+1}|s_t, a_t)$. The environment has a Markov structure: the probability of (s_{t+1}, r_{t+1}) depends only on s_t and a_t. In a finite-horizon setting with final date T, the individual's goal is to maximize the expected sum of rewards:

$$\max_{\{a_t\}} E_0 \left[\sum_{t=1}^{T} r_t \right]. \tag{1}$$

In an infinite-horizon setting, the goal is to maximize the expected sum of discounted rewards:

$$\max_{\{a_t\}} E_0 \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \right], \tag{2}$$

where $\gamma \in [0, 1)$ is a discount factor.

Economists almost always tackle a problem of this type using dynamic programming. Under this approach, we solve for the value function $V(s_t)$ – the expected sum of discounted future rewards, under the optimal policy, conditional on being in state s_t at time t. To do this, we write down the Bellman equation that $V(s_t)$ satisfies, and with the probability distribution $p(s_{t+1}, r_{t+1}|s_t, a_t)$ in hand, we solve the equation, either analytically or numerically. The solution is sometimes used for "normative" purposes – to tell the individual how he *should* act – and sometimes for "positive" purposes, to explain observed behavior.

For "positive" applications, where we are trying to explain why people behave the way they do, the dynamic programming approach raises an obvious question. It may be hard to determine the probability distribution $p(\cdot)$; and even if we have a good sense of this distribution, it may be difficult, even for professional economists, to solve the Bellman equation for the value function. How, then, would an ordinary person be able to do so? Economists

have long suggested that people act "as if" they have solved the Bellman equation – but they have not explained how this would come about. Psychologists, by contrast, have been trying to develop a more literal description of how people make decisions in dynamic settings – a framework that is rooted in the brain's actual computations. The leading such framework is the one we adopt in this paper, namely one that combines model-free and model-based learning.

We now describe the model-free learning algorithm that we use. As their name suggests, model-free algorithms tackle the problems in (1) and (2) without a "model of the world," in other words, without using any information about the probability distribution $p(\cdot)$. The model-free algorithms most commonly used by psychologists are Q-learning and SARSA. In the main part of the paper, we use Q-learning. In the Online Appendix, we show that SARSA leads to similar predictions.[7]

Q-learning works as follows. We focus on the case with the infinite-horizon goal in (2). Let $Q^*(s,a)$ be the expected sum of discounted rewards – in other words, the value of the expression

$$E_t\left[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-(t+1)}r_\tau\right] \tag{3}$$

– if the algorithm takes the action $a_t = a$ in state $s_t = s$ at time t and then continues optimally from time $t+1$ on; the asterisk indicates that, from time $t+1$ on, the optimal policy is followed. The goal of the algorithm is to estimate $Q^*(s,a)$ accurately for all possible actions a and states s so that it can select a good action in any given state.

Suppose that, at time t in state $s_t = s$, the algorithm takes an action $a_t = a$ – we describe below how this action is chosen – and that this leads to a reward r_{t+1} and state s_{t+1} at time $t+1$. Suppose also that, at time t, the algorithm's estimate of $Q^*(s,a)$ is $Q_t(s,a)$. At time $t+1$, after observing the reward r_{t+1}, the algorithm updates its estimate of $Q^*(s,a)$ from $Q_t(s,a)$ to $Q_{t+1}(s,a)$ according to

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t^{MF}[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s,a)], \tag{4}$$

where α_t^{MF} is known as the learning rate – the superscript stands for model-free – and where the term in square brackets is an important quantity known as the reward prediction error (RPE): the realized value of taking the action a – the immediate reward r_{t+1} plus a continuation value – relative to its previously anticipated value, $Q_t(s,a)$. Put simply, the

[7] Q-learning was developed by Watkins (1989) and Watkins and Dayan (1992). Sutton and Barto (2019, Ch. 6) offer a useful exposition.

updating rule in (4) says that, if, after taking the action a, the algorithm observes a better outcome than anticipated, it raises its estimate of the value of that action.

How does the algorithm choose an action a_t in state $s_t = s$ at time t? It does not necessarily choose the action with the highest estimated value of $Q^*(s, a_t)$, in other words, with the highest value of $Q_t(s, a_t)$. Rather, it chooses an action probabilistically, where the probability of choosing a given action is an increasing function of its Q value:

$$p(a_t = a|s_t = s) = \frac{\exp[\beta Q_t(s, a)]}{\sum_{a'} \exp[\beta Q_t(s, a')]}. \tag{5}$$

This probabilistic choice, known as a "softmax" specification, serves an important purpose: it encourages the algorithm to "explore," in other words, to try an action other than the one that currently has the highest Q value in order to learn more about the value of this other action. In the limit as $\beta \to \infty$, the algorithm chooses the action with the highest current Q value; in the limit as $\beta \to 0$, it chooses an action randomly. The parameter β is called the "inverse temperature" parameter, but we refer to it more simply as the exploration parameter. We discuss what exploration means in financial settings in more detail in Section 2.3.[8]

The algorithm is initialized at time 0 by setting $Q(s, a) = 0$ for all s and a. Consistent with (5), the time 0 action is chosen randomly from the set of possible actions. The process then proceeds according to equations (4) and (5). If the algorithm takes the action a in state s and this is followed by a good outcome, the value of $Q(s, a)$ goes up, making it more likely that, if the algorithm encounters state s again, it will again choose action a.

To see why equation (4) is a sensible updating rule, recall that $Q^*(s, a)$ satisfies the Bellman equation

$$Q^*(s, a) = E[r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a')|s_t = s, a_t = a], \tag{6}$$

where the expectation is taken over future possible rewards r_{t+1} and states s_{t+1} by way of the probability distribution $p(r_{t+1}, s_{t+1}|s_t, a_t)$. If we now rewrite (4) as

$$Q_{t+1}(s, a) = (1 - \alpha_t^{MF})Q_t(s, a) + \alpha_t^{MF}[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a')], \tag{7}$$

we see that the Q-learning algorithm is taking an estimate of the right-hand side of (6)

[8] Another interpretation of the probabilistic choice in (5) is that it stems from cognitive noise: due to errors in perception or cognitive processing, the algorithm does not necessarily select the action with the highest Q value. See Woodford (2020) for a review of recent research on cognitive noise.

and then updating $Q_t(s,a)$ in the direction of this estimate to an extent determined by the learning rate α_t^{MF}. Specifically, it proxies for the expected reward $E_t(r_{t+1})$ in (6) by the realized reward r_{t+1} and for $E_t[\max_{a'} Q^*(s_{t+1},a')]$ by $\max_{a'} Q_t(s_{t+1},a')$. As such, while the Q-learning algorithm differs from traditional economic approaches, it traces back to an object that is very familiar to economists, namely the Bellman equation in (6).

Computer scientists have found Q-learning to be a useful way of solving the problem in (2); under certain conditions, the Q values generated by the algorithm converge to the correct Q^* values (Watkins and Dayan, 1992). More important for our purposes, psychologists and neuroscientists are also interested in model-free algorithms like Q-learning because of evidence that they correspond to actual computations made by both animal and human brains; as noted in the previous section, many studies have found that the brain computes reward prediction errors similar to the one on the right-hand side of equation (4).[9]

When psychologists use Q-learning to explain behavior, they often allow for different learning rates for positive and negative reward prediction errors, so that

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_{t,\pm}^{MF}(\text{RPE}), \tag{8}$$

where $\alpha_{t,\pm}^{MF} = \alpha_{t,+}^{MF}$ if the reward prediction error is positive and $\alpha_{t,\pm}^{MF} = \alpha_{t,-}^{MF}$ otherwise. In what follows, we also adopt this modification.

In the basic implementation of model-free learning described above, after taking an action a in state s, the algorithm updates only the Q value for that particular action-state pair. It is natural to ask whether the algorithm can "generalize" from its experience of (s,a) to also update the Q values of other action-state pairs. We return to this below, after first introducing the financial setting that we apply the algorithm to.

2.3 A portfolio-choice setting

In Section 2.4, we lay out a model-based algorithm to complement the model-free algorithm of Section 2.2. Before we do so, it will be helpful to first describe the task that we apply

[9]Montague, Dayan, and Sejnowksi (1996) and Schultz, Dayan, and Montague (1997) made the influential observation that the activity of dopamine neurons in animal brains, as recorded in famous experiments in the laboratory of Wolfram Schultz, is well described by the reward prediction error in an important class of model-free algorithms called temporal-difference algorithms; Q-learning is a type of temporal-difference algorithm. Subsequent studies that use fMRI to study human decision-making find that neural activity in the ventral striatum correlates with the reward prediction error from model-free algorithms (McClure, Berns, and Montague, 2003; O'Doherty et al., 2003; Glascher et al., 2010; Daw et al., 2011).

both algorithms to.

We consider a simple portfolio-choice problem, namely allocating between two assets: a risk-free asset and a risky asset which we think of as the stock market. The risk-free asset earns a constant gross return R_f in each period. The gross return on the risky asset between time $t-1$ and t, $R_{m,t}$, where "m" stands for market, has a lognormal distribution

$$\begin{aligned} \log R_{m,t} &= \mu + \sigma\varepsilon_t \\ \varepsilon_t &\sim N(0,1), \text{ i.i.d.} \end{aligned} \tag{9}$$

At each time t, an investor chooses the fraction of his wealth that he allocates to the risky asset; this corresponds to the "action" in the framework of Section 2.2, so we use the notation a_t for it.[10] We construct an objective function that is realistic and also has the required form in (2). Specifically, the investor's goal is to maximize the expected log utility of wealth at some future horizon determined by his liquidity needs. Because the timing of these liquidity needs is uncertain, he does not know in advance how far away this horizon is. More precisely, at time 0, the investor enters financial markets. If, coming into time $t \geq 1$, he is still present in financial markets, then, with probability $1-\gamma$, where $\gamma \in [0,1)$, a liquidity shock arrives at time t. In that case, he exits financial markets and receives log utility from his wealth at time t. A short calculation shows that the investor's implied objective is then to solve

$$\max_{\{a_t\}} E_0\left[\sum_{t=1}^{\infty} \gamma^{t-1} \log R_{p,t}\right], \tag{10}$$

where $R_{p,t}$, the gross portfolio return between time $t-1$ and t, is given by

$$R_{p,t} = (1-a_{t-1})R_f + a_{t-1}R_{m,t}. \tag{11}$$

Comparing (2) and (10), we see that this portfolio problem maps into the framework of Section 2.2: the generic reward r_t in equation (2) now has a concrete form, namely the log portfolio return, $\log R_{p,t}$.

Given our assumptions about the returns of the two assets, we can solve the problem in (10). The solution is that, at each time t, the investor allocates the same constant fraction a^* of his wealth to the stock market, where

$$a^* = \arg\max_a E_t \log((1-a)R_f + aR_{m,t+1}). \tag{12}$$

[10] From now on, we use the terms "action" and "allocation" interchangeably.

The fact that the problem in (10) has a mathematical solution does not necessarily mean that real-world investors will be able to find their way to that solution. Many investors may have a poor sense of the statistical distribution of returns; and even if they have a good sense of it, they may not be able to compute the optimal policy or to discern it intuitively. Indeed, for many investors, the solution in (12) will *not* be intuitive, in that it involves reducing exposure to the stock market after the market has performed well and increasing exposure to the stock market after the market has performed poorly – actions that will feel unnatural to many investors.

If an investor is unable to explicitly compute the solution to the problem in (10), then, as argued in the Introduction, there is reason to think that a model-free system like Q-learning will play a role in his decision-making. As a fundamental part of human thinking, the model-free system is likely to play a role in any decision unless it is explicitly turned off. And for an investor who is unsure about the distribution of asset returns, the brain is all the more likely to assign some control to the model-free system, precisely because this system does not rely on any information about this distribution. This leads to the question at the heart of this paper: How will an investor behave if model-free Q-learning influences his actions?

How can Q-learning be applied to the above problem? In principle, we could apply equation (4) directly. However, it is natural to start with a simpler case – the case with no state dependence, so that $Q(s,a)$ is replaced by $Q(a)$. Even this simple case has rich implications that shed light on empirical facts, and so it will be our main focus. In psychological terms, removing the state dependence can be thought of as a simplification on the part of the investor. Indeed, neuroscience research has argued that, to speed up learning, the brain does try to simplify the state structure when implementing its learning algorithms (Collins, 2018).[11] While, in the main body of the paper, we put state dependence aside, in the Online Appendix, we re-introduce it and confirm that the key properties of the framework continue to hold.

As in Section 2.2, then, let $Q^*(a)$ be the expected sum of discounted rewards – in other words, the value of

$$E_t\left[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-(t+1)}\log R_{p,\tau}\right]$$

– if the investor chooses the allocation a at time t and then continues optimally from the next period on. Suppose that, at time t, the investor chooses the allocation a and observes the

[11] It is tempting to justify the removal of the state dependence by saying that, since the risky asset returns are i.i.d., the allocation problem has the same form at each time and so there is no state dependence. However, we cannot use this argument because the model-free system does not know that the returns are i.i.d.; by its nature, it does not have a model of the environment.

reward – the log portfolio return, $\log R_{p,t+1}$ – at time $t+1$. He then updates his model-free estimate of $Q^*(a)$ from $Q_t^{MF}(a)$ to $Q_{t+1}^{MF}(a)$ according to

$$Q_{t+1}^{MF}(a) = Q_t^{MF}(a) + \alpha_{t,\pm}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)]. \tag{13}$$

At any time t, he chooses his allocation a_t probabilistically, according to

$$p(a_t = a) = \frac{\exp[\beta Q_t^{MF}(a)]}{\sum_{a'} \exp[\beta Q_t^{MF}(a')]}. \tag{14}$$

Put simply, if the investor chooses an allocation a and then experiences a good portfolio return, this tends to increase the Q value of that allocation and makes it more likely that he will choose that allocation again in the future.

The exploration embedded in (14) is central to the model-free algorithm and to the way psychologists think about human behavior. By contrast, the term is less common in economics and finance. Nonetheless, many actions in financial settings can be thought of as forms of exploration – for example, any time an individual tries a strategy that is new to him, such as investing in a stock in a different industry or foreign country, or in an entirely new asset class. In our context, with one risk-free and one risky asset, exploration can be thought of as the investor choosing a different allocation to the stock market than before in order to learn more about the value of doing so.[12]

Given our assumption about the distribution of stock market returns, we can compute the exact value of $Q^*(a)$ for any allocation a. We record it here because we will use it in the next section. It is given by

$$Q^*(a) = E\log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E\log((1-a^*)R_f + a^* R_{m,t+1}), \tag{15}$$

where a^* is defined in (12).

In the basic model-free algorithm in (13), after taking action $a_t = a$ at time t, only the Q value of action a is updated. It is natural to ask whether the algorithm can generalize from its experience of taking the action a in order to also update the Q values of other actions. Computer scientists have studied model-free generalization (Sutton and Barto, 2019, Chs. 9-13). As important for our purposes, research in psychology suggests that the human model-free system engages in generalization (Shepard, 1987). We therefore incorporate generalization into our framework.

[12] As noted in Section 2.2, another possible foundation for the probabilistic choice in (14), one that may be relevant in financial settings, is cognitive noise.

Given that we are working with the model-free system, it is important that the generalization we consider does not use any information about the structure of the allocation problem. We adopt a simple form of generalization based on the notion of similarity: after choosing an allocation and observing the subsequent portfolio return, the algorithm updates the Q values of all allocations, but particularly those that are similar to the chosen allocation. We implement this as follows. After choosing allocation a at time t and observing the outcome at time $t+1$, the algorithm updates the values of all allocations according to

$$Q_{t+1}^{MF}(\widehat{a}) = Q_t^{MF}(\widehat{a}) + \alpha_{t,\pm}^{MF}\kappa(\widehat{a})[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)], \tag{16}$$

where

$$\kappa(\widehat{a}) = \exp(-\frac{(\widehat{a}-a)^2}{2b^2}). \tag{17}$$

In words, after observing the reward prediction error for action a and updating the Q value of that action, the algorithm uses the *same* reward prediction error to also update the values of all other actions. However, for an action $\widehat{a}$ that differs from a, it uses a lower learning rate $\alpha_{t,\pm}^{MF}\kappa(\widehat{a})$, one that is all the lower, the more different $\widehat{a}$ is from a, to an extent determined by the Gaussian function in (17).[13]

We will consider a range of values of b, but for our baseline analysis, we set $b = 0.0577$, which has a simple interpretation: for this b, the Gaussian function in (17), normalized to form a probability distribution, has the same standard deviation as a uniform distribution with width 0.2 – for example, the uniform distribution that ranges from $a-10\%$ to $a+10\%$. For this b, then, the model-free algorithm generalizes primarily to nearby allocations, those within ten percentage points of the chosen allocation. We later examine the sensitivity of our results to the value of b.[14]

We emphasize that the Q-learning algorithm above, with or without generalization, does not use any information about the distribution of risky asset returns in (9): by its model-free nature, it does not have a model of the environment. More broadly, the algorithm has no

[13]Our generalization algorithm is consistent with research in psychology which identifies similarity as an important driver of generalization (Shepard, 1987). It is also used in computer science, where it is known as interpolation-based Q-learning (Szepesvari, 2010, Ch. 3.3.2). Computer scientists also use more sophisticated forms of generalization such as function approximation with polynomial, Fourier, or Gaussian basis functions (Sutton and Barto, 2019, Ch. 9). We have also implemented this more complex generalization and obtain similar results.

[14]One interpretation of our generalization algorithm is that the model-free system uses a *small* amount of "model" information, namely that similar allocations lead to similar portfolio returns; as such, after observing the outcome of a 70% allocation, the system updates the Q value of an 80% allocation more than that of a 20% allocation. An alternative interpretation – a strictly model-free interpretation that uses no information about the structure of the task – is that the generalization is based simply on numerical similarity: the number 70 is closer to 80 than to 20.

idea what a "risk-free asset" or the "stock market" are. It is simply choosing an action – some combination of these unfamiliar objects – seeing what reward it delivers, and then updating the values of the chosen action and of actions similar to it. While the model-free system may appear uninformed, the fact that it uses so little information about the problem at hand is precisely what makes it powerful, in general: it can be applied in almost any setting. Moreover, its implications will turn out to be helpful for thinking about a range of facts in finance.

2.4 Model-based learning

Current research in psychology uses a framework in which decisions are guided by both model-free and model-based learning. Model-based systems, as their name indicates, build a model of the environment, which, more concretely, means a probability distribution of future outcomes – for example, in our setting, a probability distribution of stock market returns. There are various possible model-based systems. Which one should we choose? Our goal in this paper is to see if algorithms commonly used by psychologists can explain behavior in economic settings. We therefore take as our model-based system one that, like the model-free system of Section 2.2, is based on an algorithm that is used extensively by psychologists and is supported by neural evidence from decision-making experiments.

In our model-based system, an investor learns the distribution of stock market returns over time by observing realized market returns. At each date, he updates the probabilities of different returns using prediction errors analogous to the reward prediction errors of Section 2.2. Specifically, suppose that the investor observes a stock market return $R_{m,t+1} = R$ at time $t+1$ and that, at time t, before observing the return, the prior probability he assigned to it occurring was $p_t(R_m = R)$. At time $t+1$, he updates the probability of this return as

$$p_{t+1}(R_m = R) = p_t(R_m = R) + \alpha_t^{MB}[1 - p_t(R_m = R)], \tag{18}$$

where α_t^{MB} is the model-based learning rate that applies from time t to time $t+1$. The term $1 - p_t(R_m = R)$ is a prediction error: the investor's prior estimate of the probability of the return equaling R was $p_t(R_m = R)$; when the return is realized, the probability of it equaling R is 1. After this update, the investor scales the probabilities of all other returns down by the same proportional factor so that the sum of all return probabilities continues to equal one. Since we are working with a continuous return distribution, we can assume that each return that is realized is one that has not been realized before. As such, $p_t(R_m = R) = 0$,

which simplifies (18) to

$$p_{t+1}(R_m = R) = \alpha_t^{MB}.$$

To illustrate this process, suppose that the investor observes four stock market returns in sequence: R_1, R_2, R_3, and R_4, at dates 1, 2, 3, and 4, respectively. The four rows below show the investor's perceived probability distribution of stock market returns at dates 1, 2, 3, and 4, in the case where the learning rate is constant over time, so that $\alpha_t^{MB} = \alpha$ for all t. In this notation, a comma separates a return from its perceived probability, while semicolons separate the different returns:

$$\begin{aligned}
&(R_1, 1)\\
&(R_1, 1-\alpha; R_2, \alpha)\\
&(R_1, (1-\alpha)^2; R_2, \alpha(1-\alpha); R_3, \alpha)\\
&(R_1, (1-\alpha)^3; R_2, \alpha(1-\alpha)^2; R_3, \alpha(1-\alpha); R_4, \alpha).
\end{aligned} \tag{19}$$

The above approach is motivated by research in decision neuroscience that adopts a similar model-based system (Glascher et al., 2010; Lee, Shimojo, and O'Doherty, 2014; Dunne et al., 2016). Just as there is evidence that the brain encodes reward prediction errors, so there is evidence that it encodes prediction errors analogous to the one in square brackets in (18) (Glascher et al., 2010).[15]

We noted in Section 2.2 that, when they implement model-free learning, psychologists allow for different model-free learning rates, α_+^{MF} and α_-^{MF}, for positive and negative reward prediction errors, respectively. We extend the model-based algorithm in a similar way, allowing for different model-based learning rates, α_+^{MB} and α_-^{MB}, for positive and negative net stock market returns, respectively. Specifically, following the gross return $R_{m,t+1} = R$,

$$p_{t+1}(R_m = R) = \alpha_{t,+}^{MB} \text{ for } R \geq 1, \tag{20}$$

with the probabilities of all other returns being scaled down by $1 - \alpha_{t,+}^{MB}$, and

$$p_{t+1}(R_m = R) = \alpha_{t,-}^{MB} \text{ for } R < 1, \tag{21}$$

with the probabilities of all other returns being scaled down by $1 - \alpha_{t,-}^{MB}$.

[15]While our model-based algorithm is inspired by research in psychology, it is also similar to an existing economic framework, namely adaptive learning (Evans and Honkapohja, 2012). As such, from the perspective of economics, the novel elements of our framework are the model-free system and its interaction with its model-based counterpart.

With this perceived return distribution in hand, how does the investor come up with a model-based estimate of $Q^*(a)$, the value of choosing an allocation a on some date and then continuing optimally thereafter? We again follow an approach taken by experimental studies in decision neuroscience (Glascher at al., 2010). We assume that, for any allocation a, the individual computes his time t model-based estimate of $Q^*(a)$, denoted $Q_t^{MB}(a)$, by taking the correct form of $Q^*(a)$ in equation (15) and applying it for his *perceived* time t return distribution:

$$Q_t^{MB}(a) = E_t^p \log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E_t^p \log((1-a_t^*)R_f + a_t^* R_{m,t+1}), \tag{22}$$

where

$$a_t^* = \arg\max_a E_t^p \log((1-a)R_f + aR_{m,t+1}) \tag{23}$$

and where (22) differs from (15) only in that the expectation E under the correct distribution has been replaced by the expectation E_t^p under the investor's perceived distribution at time t.

The Daw et al. (2011) experiment discussed in Section 2.1 illustrates a tension between the model-free and model-based systems. If, in that experiment, an individual chooses A and then E and is rewarded, the model-free system wants to repeat action A in the next round, while the model-based system, recognizing that B is more likely to lead to E, wants to choose B. The same tension is present in our financial market setting. If the investor starts with a low allocation to the stock market and the market then posts a high return, the model-free system wants to stick with a low allocation because this action was "reinforced": it was followed by a positive reward prediction error. By contrast, the model-based system wants to increase the investor's allocation to the stock market: it now perceives a more attractive distribution of market returns and wants more exposure to it. We explore the implications of this tension in Section 3.

The model-free and model-based systems are not the only learning algorithms the brain uses. Another important class of algorithms are "observational learning" algorithms which learn by observing the actions and outcomes of other people (Charpentier and O'Doherty, 2018). We focus on the model-free and model-based algorithms because they have received the most attention from psychologists and because they likely "span" other algorithms: these other learning systems tend to generate predictions that lie somewhere between those of the model-free and model-based systems.

2.5 A hybrid framework

An influential framework in psychology posits that people make decisions using a combination of model-free and model-based systems (Daw, Niv, and Dayan, 2005; Glascher et al., 2010; Daw et al., 2011). Specifically, it proposes that, at each time t, and for each possible action a, an individual computes a "hybrid" estimate of $Q^*(a)$, denoted $Q_t^{HYB}(a)$, that is a weighted average of the model-free and model-based Q values:

$$Q_t^{HYB}(a) = (1-w)Q_t^{MF}(a) + wQ_t^{MB}(a), \tag{24}$$

where w is the weight on the model-based system. He then chooses an action using the softmax approach, now applied to the hybrid Q values:

$$p(a_t = a) = \frac{\exp[\beta Q_t^{HYB}(a)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(a')]}. \tag{25}$$

In this paper, we focus on the case where w is constant over time, as this already leads to a rich set of properties and applications. Nonetheless, a well-known hypothesis in psychology is that w varies over time: at each moment of time, the brain puts more weight on the system it deems more "reliable" at that point (Daw, Niv, and Dayan, 2005). In Section A of the Online Appendix, we formalize and explore this idea using an implementation proposed by researchers in decision neuroscience in which a system's reliability is measured by the absolute magnitude of its prediction errors: if the model-free reward prediction errors have been large in absolute magnitude, the brain deems the model-free system to be less reliable and raises w, thereby allocating more control to the model-based system. We discuss the implications of this idea in Section 5 and in the Online Appendix.

The model-free and model-based systems differ most fundamentally in how they estimate the value of an action: one system uses a model of the environment, while the other does not. However, there is another difference between them: the model-free system learns only from experienced rewards, while the model-based system can learn from all observed rewards. In our setting, the investor enters financial markets at time 0. Time 0 is therefore the moment at which he starts experiencing returns and hence the moment at which the model-free system begins learning. However, before he makes a decision at time 0, the investor can look at historical charts and observe earlier stock market returns, which the model-based system can then learn from. To incorporate this, we extend the timeline of our framework so that it starts not at time 0 but L dates earlier, at time $t = -L$. While the model-free system starts operating at time 0, the model-based system starts operating at time $-L$: it observes

the L stock market returns prior to time 0, $\{R_{-L+1}, \ldots, R_0\}$; uses these to form a perceived distribution of market returns as in (20) and (21); and then computes model-based Q values by way of that distribution, as in (22).[16]

3 Properties of Model-free and Model-based Systems

We begin this section with an example that illustrates the mechanics of the model-free and model-based systems. We then analyze some key properties of the framework. Our focus is on how the allocations recommended by the model-free and model-based systems depend on past stock market returns. We also examine the dispersion and variability in investor allocations that these systems generate. In Section 4, we build on these properties to account for several facts about investor behavior.

We use the timeline previewed at the end of the previous section. There are $L + T + 1$ dates, $t = -L, \ldots, -1, 0, 1, \ldots, T$. Investors begin actively participating in financial markets at time 0. Their model-free systems therefore start operating only at time 0, while their model-based systems operate over the full time range, starting from $t = -L$. We think of each time period as one year and set $L = T = 30$. Before they start investing at time 0, then, people have access to 30 years of prior data going back to $t = -30$. We then track their allocation decisions over the next 30 years, from $t = 0$ to $t = 30$.[17,18]

The four learning rates – α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} – play an important role in our framework. How should they be set? If we were taking a normative perspective – if we wanted to use the algorithms of Section 2 to solve the problem in (10) as efficiently as possible – the answer would be to use learning rates that decline over time. Specifically, the

[16]Our implementation here is consistent with evidence from decision neuroscience. Dunne et al. (2016) conduct an experiment in which participants actively experience slot machines that deliver a stochastic reward, but also passively observe other people playing the slot machines. fMRI measurements show that, as in many other studies, the model-free reward prediction error for the experienced trials is encoded in the ventral striatum. However, for the trials that are merely observational, the model-free RPE is *not* encoded in the striatum, suggesting that the model-free system is not engaged. As Dunne et al. (2016) write, "It may be that the lack of experienced reward during observational learning prevents engagement of a model-free learning mechanism that relies on the receipt of reinforcement."

[17]One interpretation of our annual implementation is that, as argued by Benartzi and Thaler (1995), investors pay particular attention to their portfolios once a year – at tax time, or when they receive their end-of-year brokerage statements. Another interpretation is that it is an approximation of a higher-frequency implementation. Later in this section, we explain how our results are affected by the choice of frequency.

[18]Since our setting has an infinite horizon, investors continue to participate in financial markets beyond date T. Date T is simply the date at which we stop tracking their allocation decisions.

time t model-based learning rates in (20) and (21) would be[19]

$$\alpha_{t,+}^{MB} = \alpha_{t,-}^{MB} = 1/(t+1), \tag{26}$$

as these lead investors to equally weight all past returns, consistent with the i.i.d. return assumption. Similarly, Watkins and Dayan (1992) show that, for Q-learning to converge to the correct Q^* values, declining model-free learning rates are needed that, for each action a, satisfy

$$\sum_{t=0}^{\infty} \alpha_{t,\pm}^{MF} 1_{\{a_t=a\}} = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} (\alpha_{t,\pm}^{MF})^2 1_{\{a_t=a\}} < \infty, \tag{27}$$

where the indicator function identifies periods where the algorithm is taking action a.

In this paper, however, we are taking a "positive" perspective – our goal is to explain observed behavior. What matters for our purposes is therefore not the learning rates people should use, but rather the learning rates they actually use. Psychology research does not offer definitive guidance on people's learning rates, but most studies of actual decision-making use learning rates that are constant over time (Glascher et al., 2010). For this reason, we focus on constant learning rates. To start, we give all investors the same constant learning rates. Later, we allow for dispersion in these rates across investors.

3.1 An example

To show how the model-free and model-based systems work, we start with an example. Throughout the paper, we use the same baseline parameter values, in part for consistency and in part to show that a fixed set of parameter values can account for a range of observed facts. We consider an investor who is exposed to a sequence of stock market returns from $t = -L$ to $t = T$, where $L = T = 30$. The returns are simulated from the distribution in (9) with $\mu = 0.01$ and $\sigma = 0.2$; these values provide an approximate fit to historical annual U.S. stock market data. We set the investor's learning rates to $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, the exploration parameter β to 30, the discount factor γ to 0.97 – this corresponds to an expected investment horizon of 33 years – and the degree of generalization b to 0.0577.[20] At each date, we allow the investor to choose his stock market allocation a_t from one of 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$. We later examine how our results depend on the values of all the key parameters.

[19]Equation (26) assumes $L = 0$. For $L > 0$, the learning rates would be $\alpha_{t,+}^{MB} = \alpha_{t,-}^{MB} = 1/(L+t+1)$.

[20]In simulations, we find that for $\beta = 30$, an investor using the hybrid system chooses the allocation with the highest Q value approximately half the time, which represents a moderate degree of exploration.

In our framework, decisions are based on hybrid Q values that combine the influences of the model-free and model-based systems. To clearly illustrate the mechanics of each system, we start by considering two simpler cases: one where the investor uses only the model-free system to make decisions, and one where he uses only the model-based system.

Table 1 shows the model-free Q values, Q^{MF}, based on equations (14), (16), and (17) (upper panel) and the model-based Q values, Q^{MB}, based on equations (22) and (23) (lower panel) that the investor assigns to the 11 allocation strategies on his first six dates of participation in financial markets, namely $t = 0$, 1, 2, 3, 4, and 5. The rows labeled "net market return" show the net return of the stock market at each date. In each column, the number in bold corresponds to the action that was taken in the previous period; for example, the number -0.065 in bold at date 1 in the upper table indicates that the investor chose a 70% allocation at date 0.[21]

Consider the upper panel of Table 1. The model-free system begins operating at time 0. At that time, then, it assigns a Q value of zero to all the allocations. It then randomly selects the allocation 70%. The net stock market return at time 1 is negative, which means that the investor's net portfolio return and reward prediction error are also negative. The time-1 Q value for the 70% allocation therefore falls below zero. As per equations (16) and (17), the algorithm also engages in some generalization: since a 60% allocation and an 80% allocation are similar to a 70% allocation, their Q values also fall, albeit to a lesser extent. The Q values of more distant allocations are unaffected, at least to three decimal places.

At time 1, the investor chooses the allocation 30%. The time-2 market return is positive; the investor therefore earns a positive net portfolio return and the time-2 Q value of the 30% allocation goes up, as do, to a lesser extent, the Q values of the similar allocations 20% and 40%. At time 2, the investor chooses the allocation 100%. While the market falls slightly at time 3, the time-3 Q value of the 100% allocation goes up by a small amount because the reward prediction error is slightly positive. At dates 3 and 4, the investor chooses allocations of 30% and 40%, respectively, and updates the values of these allocations and their close neighbors based on the prediction errors they lead to at dates 4 and 5.

[21] In the case where decisions are determined by the model-based system alone, we assume that the investor still chooses actions probabilistically, in a manner analogous to that in (14). In our setting, for the model-based system, this probabilistic choice does not offer the usual exploration benefits: in each period, the investor learns the same thing about the distribution of stock market returns regardless of which allocation he chooses. We keep the probabilistic choice to allow for a more direct comparison with the model-free system – but also because, if, as suggested earlier, this stochastic choice stems in part from cognitive noise, it will be relevant for model-based learning too. For these reasons, whenever we consider the model-based system in isolation, we will allow for probabilistic choice, unless otherwise specified.

Table 1. Model-free and model-based Q values. The upper panel reports model-free Q values for 11 stock market allocations from $t = 0$ to $t = 5$. The lower panel reports model-based Q values for the 11 allocations for the same six dates. The rows labeled "net market return" report the net stock market return at each date. Boldface type indicates the allocation that was taken in the previous period. We set $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

MODEL-FREE						
date	0	1	2	3	4	5
net market return		-17.4%	18.3%	-1.3%	12.8%	-16.6%
0%	0	0	0	0	0	0
10%	0	0	0	0	0	0
20%	0	0	0.006	0.006	0.01	0.01
30%	0	0	**0.027**	0.027	**0.045**	0.041
40%	0	0	0.006	0.006	0.01	**-0.007**
50%	0	0	0	0	0	-0.004
60%	0	-0.015	-0.015	-0.015	-0.015	-0.015
70%	0	**-0.065**	-0.065	-0.065	-0.065	-0.065
80%	0	-0.015	-0.015	-0.014	-0.014	-0.014
90%	0	0	0	0.001	0.001	0.001
100%	0	0	0	**0.006**	0.006	0.006

MODEL-BASED						
date	0	1	2	3	4	5
net market return		-17.4%	18.3%	-1.3%	12.8%	-16.6%
0%	0.72	0	**1.352**	0.464	2.179	0
10%	0.723	-0.007	1.357	0.466	**2.187**	-0.005
20%	0.726	-0.015	1.362	0.468	2.194	-0.01
30%	0.729	-0.022	1.367	0.47	2.201	-0.015
40%	0.731	-0.03	1.372	0.472	2.208	**-0.02**
50%	0.733	**-0.039**	1.376	0.473	2.215	-0.026
60%	0.736	-0.047	1.38	0.475	2.222	-0.031
70%	0.737	-0.056	1.384	0.476	2.228	-0.037
80%	0.739	-0.065	1.387	0.477	2.234	-0.044
90%	0.741	-0.075	1.39	**0.478**	2.241	-0.05
100%	0.742	-0.085	1.393	0.479	2.247	-0.057

The lower panel shows that the Q values generated by the model-based system are quite different. By time 0, the model-based system has already been operating for 30 periods and so already has well-developed Q values for each of the 11 allocation strategies. In the periods immediately preceding time 0, the simulated stock market returns are somewhat positive; higher allocations to the stock market therefore have higher Q values at time 0. At time 1, the stock market return is poor, so all Q values fall, but those of riskier allocations do so more: the negative stock market return at time 1 makes the investor's perceived distribution of stock market returns less appealing; this has a larger impact on strategies that allocate more to the stock market. At time 2, the stock market return is positive, so all Q values go up, but those of the riskier allocations do so more.

Table 1 makes clear a key difference between the model-free and model-based systems: while, at each time, the model-based system updates the Q values of all the allocations, the model-free system primarily updates only the Q values of the most recently chosen allocation and those of its nearest neighbors. The reason is that it is model-free: it knows nothing about the structure of the problem and therefore cannot make a strong inference, after seeing the outcome of a 70% allocation, about the value of a 20% allocation.

3.2 Dependence on past market returns

We now analyze a property of our framework that is central to the applications in Section 4, namely, how the stock market allocations recommended by the model-free and model-based systems depend on past stock market returns. We find that the model-free system generates a rich set of intuitions and implications, some of which are quite distinct from those associated with model-based systems.

To study this, we take $300,000$ investors and expose each of them to a different sequence of simulated stock market returns from $t = -L$ to $t = T$. We then take investors' stock market allocations a_T at time T, regress them on the past 30 annual stock market returns $\{R_{m,T}, R_{m,T-1},\ldots,R_{m,T-29}\}$ the investors have been exposed to, and record the coefficients. We do this for three cases, namely those where investor allocations are determined by the model-free system alone; by the model-based system alone; and by the hybrid system. For all investors, as before, we set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. For ease of interpretation, we turn off generalization for now, so that $b = 0$.[22] Finally, we set $w = 0.5$, so that the hybrid system puts equal weight on the model-free and

[22]We use "$b = 0$" as shorthand for model-free learning without generalization. When $b = 0$, we compute model-free Q values using equation (13) rather than equations (16)-(17), although the latter equations give the same result as $b \to 0$.

model-based systems. We later look at how changing the values of key model parameters affects the results.[23]

Figure 2 presents the results. The solid line plots the coefficients on past returns in the above regression when allocations are determined by the model-based system. As we move from left to right, the line plots the coefficients on more distant past returns: the point on the horizontal axis that marks j years in the past corresponds to the coefficient on $R_{m,T+1-j}$. The two other lines plot the coefficients for the model-free and hybrid systems.

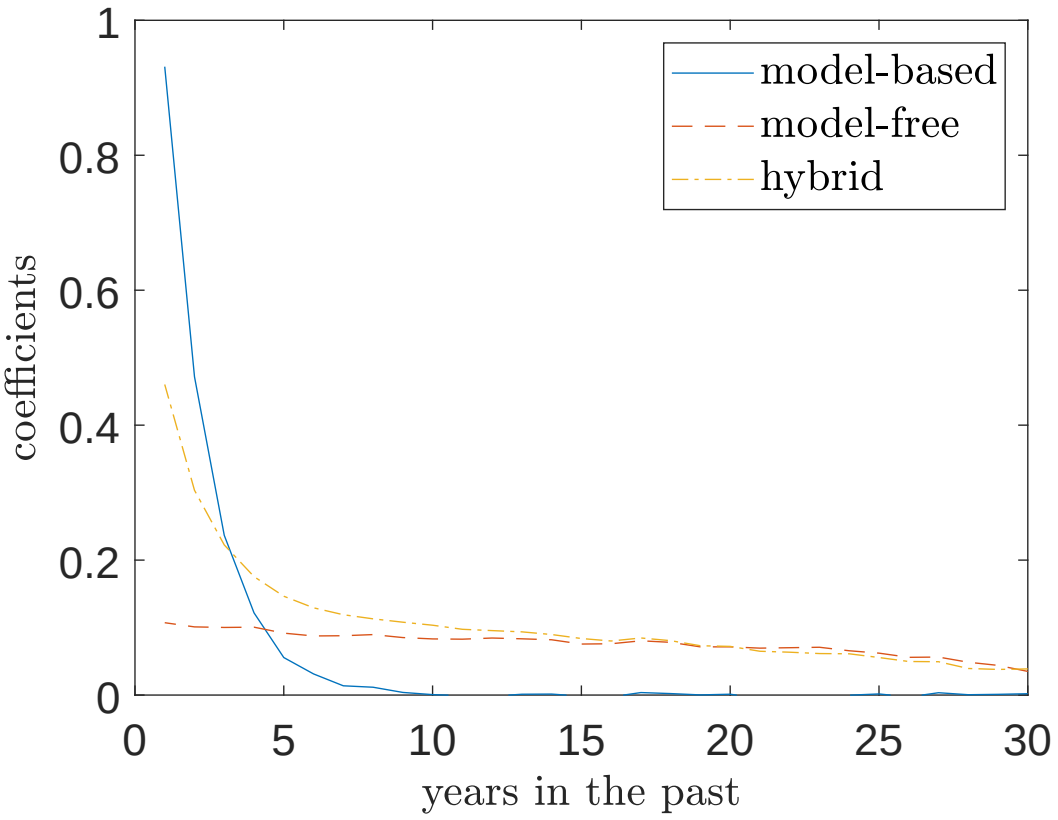

Figure 2. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system. The point on the horizontal axis that marks j years in the past corresponds to the coefficient on $R_{m,T+1-j}$. There are 300,000 investors. We set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, $w = 0.5$, and $b = 0$, so that there is no generalization.

[23]The goal function in (10) is motivated in part by the idea that, due to liquidity shocks, some investors drop out of financial markets over time. In our calculations, we do not explicitly track which investors drop out. This is because the shocks are random: they do not depend on investors' prior allocations or past returns. As such, investor exits do not affect the properties or predictions that we document.

The figure shows that, for both the model-free and model-based systems, the time T stock market allocations depend positively on past returns, and more so on recent past returns: the coefficients on past returns decline, the more distant the past return. Importantly, the decline is much more gradual for the model-free system, a property that will play a key role in some of our applications. Given that the hybrid system combines the model-free and model-based systems, it is natural that the line for the hybrid system is, approximately, a mix of the model-free and model-based lines.

We now discuss these findings. First, we explain why the allocations recommended by the model-free and model-based systems depend positively on past returns. The answer is clear in the case of the model-based system. Following a good stock market return, an investor's perceived distribution of market returns assigns a higher probability to good returns and a lower probability to bad returns. This raises the model-based Q values of all stock market allocations, but particularly those of high allocations, making it more likely that the investor will choose a high allocation going forward.

The intuition in the case of the model-free system is different and, to our knowledge, new to financial economics. If the investor chooses a 20% stock market allocation and the market posts a high return, this "reinforces" the action of choosing a 20% allocation: the positive reward prediction error raises the Q value of this allocation, making it more likely that the investor will choose it again in the future. Similarly, if he chooses an 80% allocation and the market posts a high return, this reinforces the 80% allocation. In one case, then, a high market return leads the investor to choose a low allocation; in the other, it leads him to choose a high allocation. Why then, on average, does a high market return lead to a higher allocation, as shown by the dashed line in Figure 2? The reason is that the reinforcement is stronger in the case of the 80% allocation: a high stock market return leads to a larger reward prediction error when the investor's prior allocation is 80% than when it is 20%. As such, the net effect of a good stock market return, after averaging over the possible prior allocations, is to lead the investor to choose a high stock market allocation.

We now explain why the weights that the two systems put on past market returns decline as we go further into the past. In the case of the model-based system, this is because, when this system updates its perceived return distribution after seeing a new stock market return, it scales down the probabilities of earlier returns, reducing their importance. Intuitively, by using a constant learning rate, the investor is acting as if the environment is non-stationary; as such, he puts greater weight on recent returns. The top graph in Figure 3 shows how the time T allocation recommended by the model-based system depends on past stock market returns for four different values of the learning rates α_+^{MB} and α_-^{MB}, namely 0.05, 0.1, 0.2,

and 0.5. The graph shows that, regardless of the learning rate, the allocation puts weights on past returns that are positive and that decline the further back we go into the past, with the decline being more pronounced for higher learning rates.

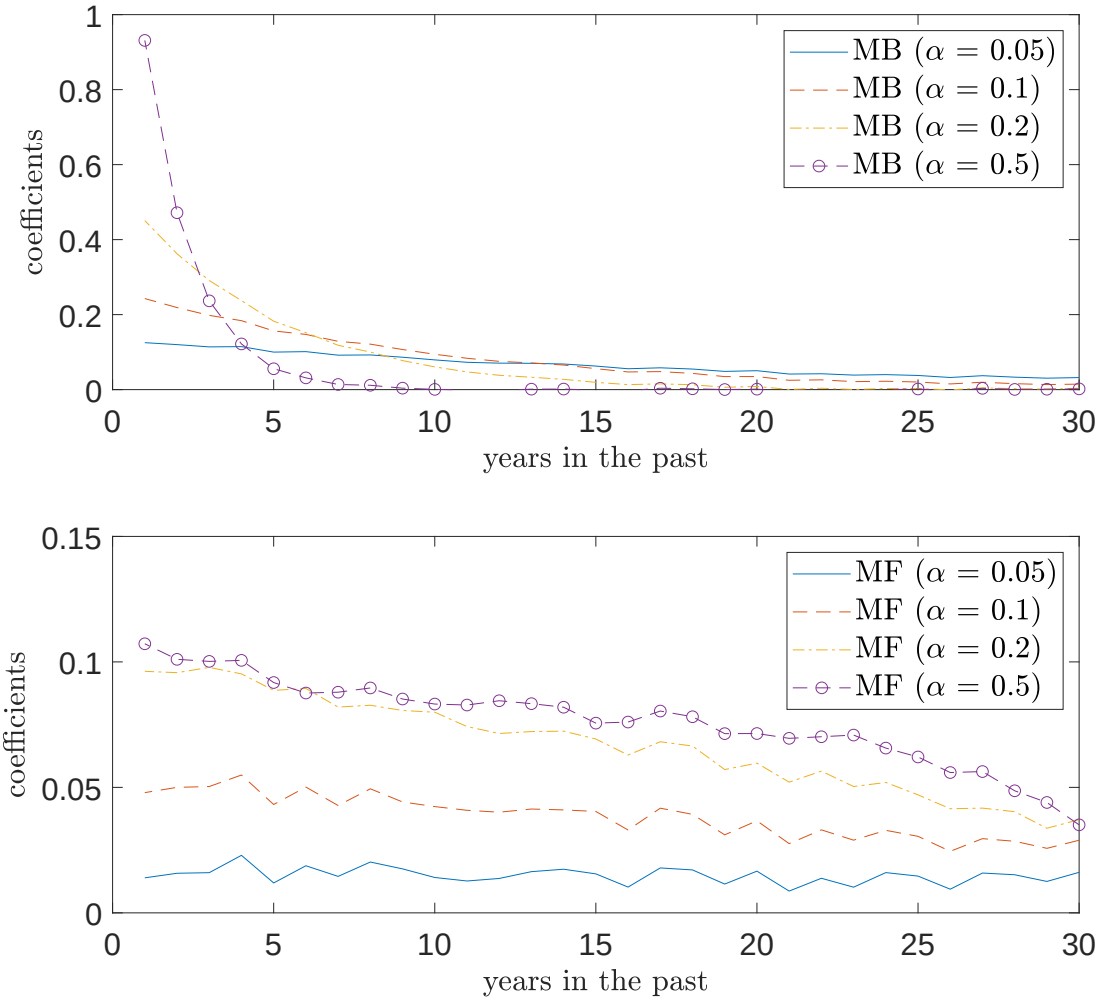

Figure 3. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to. The top graph plots the coefficients for the model-based system for four values of the learning rates α_+^{MB} and α_-^{MB}, namely 0.05, 0.1, 0.2, and 0.5. The point on the horizontal axis that marks j years in the past corresponds to the coefficient on $R_{m,T+1-j}$. The bottom graph plots the coefficients for the model-free system for four values of the learning rates α_+^{MF} and α_-^{MF}, namely 0.05, 0.1, 0.2, and 0.5. There are 300,000 investors. We set $L = T = 30$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

Figure 2 shows that, for the model-free system, the weights on past returns again decline as we go further into the past, but much more gradually. Why is this? When the model-free system updates the Q value of an action, this tends to downweight the influence of past returns on this Q value, relative to the most recent return. However, this effect passes through to allocation choice in a much more gradual way than for the model-based system because, at each time, the model-free system primarily updates only one Q value, namely that of the most recently-chosen action; as such it takes much longer for past returns to lose their influence on the investor's allocation.[24] The bottom graph in Figure 3, which plots the relationship between the model-free allocation and past returns for four different values of the learning rates α_+^{MF} and α_-^{MF}, shows that the model-free allocation typically puts positive and declining weights on past returns, with the decline being more pronounced for higher learning rates.

The graphs in Figure 4 show how the relationship between investors' time T model-free allocations and past stock market returns changes as we vary one of the model parameters while keeping the others at their benchmark levels. Across the four graphs, we vary the degree of generalization, the degree of exploration, the discount factor, and the number of allocation choices. Changing these parameters would have little effect on model-*based* allocations. However, Figure 4 shows that it has significant impact on model-free allocations. While for many parameter values, including those used in Figures 2 and 3, the model-free allocation puts more weight on recent than on distant past returns, Figure 4 shows that, for some parameter values, it can put more weight on distant than on recent past returns. Moreover, the figure shows the conditions under which this happens – for example, for higher degrees of generalization. We explain the full intuition for the patterns in Figure 4 in Online Appendix B.[25]

[24]For an example, consider the upper panel of Table 1. At time 4, the model-free system updates the Q value of the 30% allocation. However, the Q value of a 70% allocation is not significantly updated at this time, and so it depends as strongly as before on the time 1 stock market return. As such, for the model-free system, the time 1 and time 4 stock market returns exert a similar degree of influence on the investor's allocation at time 4.

[25]The results in Figures 2 to 4 are for an annual-frequency implementation of our framework. We have studied the effect of changing the frequency. If we fix the learning rates $\alpha_\pm^{MB}$ and $\alpha_\pm^{MF}$ but switch to a semi-annual, quarterly, or monthly implementation, this has a significant effect on the model-based allocation – it depends all the more on recent returns – but a much smaller impact on the model-free allocation. As such, implementing the framework at a higher frequency creates a larger wedge between the two systems.

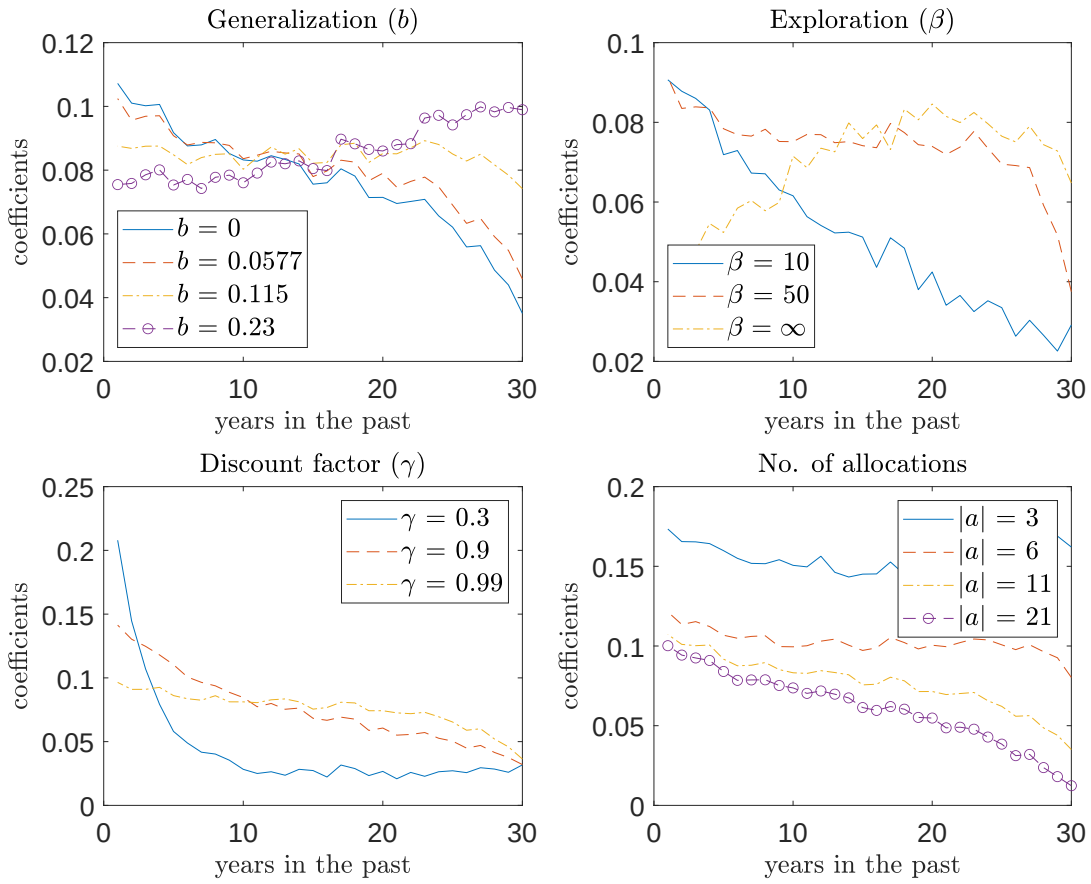

Figure 4. For different sets of parameter values, we run a regression of investors' allocations to the stock market a_T at time T under the model-free system on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients. The lines in the top-left, top-right, bottom-left, and bottom-right graphs correspond, respectively, to four values of the generalization parameter b, namely 0, 0.0577, 0.115, and 0.23; to three values of the exploration parameter β, namely 10, 50, and ∞, which corresponds to no exploration; to three values of the discount factor γ, namely 0.3, 0.9, and 0.99; and to different numbers of allocation choices, namely 3, 6, 11, and 21. There are 300,000 investors. The benchmark parameter values are $L = T = 30$, $\alpha_{\pm}^{MF} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

While the model-free algorithm is simple to state – it is summarized in equation (16) – it is difficult to derive analytical results about its predictions. Nonetheless, for certain special cases, we *are* able to derive such results, which are precisely about the dependence of model-free allocations on past market returns and which, to our knowledge, are the first results of their kind. We present these results and their proofs in Theorems 1 to 4 of Online Appendix C. Specifically, we show that, under some conditions, as $t \to \infty$, the sensitivity of the expected allocation at time t to the market return k periods earlier, $R_{m,t-k} = R$, is given by

$$\frac{\partial E(a_t)}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^3}\left(\frac{R^\beta+1-\alpha R^\beta}{R^\beta+1}\right)^k \tag{28}$$

for the model-free allocation, and by

$$\frac{\partial E(a_t)}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{\beta-1}}{(R^\beta+1)^2}(1-\alpha)^k \tag{29}$$

for the model-based allocation, where α is the constant learning rate for both systems. These results are consistent with the patterns in Figure 2. Expressions (28) and (29) both decline monotonically as k increases. Moreover, the model-free coefficient in (28) is lower than the model-based coefficient in (29) for low values of k, but higher than the model-based coefficient for high values of k. This provides an analytical foundation for the property we have emphasized in this section, namely that, relative to the model-based system, the model-free system puts significantly more weight on distant past returns.

3.3 Dependence on past allocations and portfolio returns

In the previous section, we studied the dependence of the model-free and model-based allocations on past market returns. We focused on the market return because it is the exogenous shock in our framework and because the dependence on market returns is central to our applications in Section 4. Nonetheless, it is natural to ask whether other variables – the investor's past allocations or portfolio returns – also have predictive power for today's allocation.

In the case of the model-based system, the answer is negative: past market returns are the lone predictors of today's model-based allocation; past allocations and portfolio returns have no additional predictive power. By contrast, past allocations and portfolio returns have substantial predictive power for the model-*free* allocation – indeed, they are the primary drivers of this allocation. Past *market* returns affect the model-free allocation indirectly, by way of these other variables: they affect portfolio returns which then reinforce allocations. In this section, we examine the dependence of the model-free allocation on past allocations

and portfolio returns.

A large part of the predictive power of past allocations and portfolio returns for the model-free allocation comes from just one lag of prior data. In the case where decisions are made by the model-free system, we run a regression in our simulated data of the time T allocation a_T on the prior allocation a_{T-1}, on the most recent net portfolio return $r_{p,T} \equiv R_{p,T} - 1$ – in this section, we work with the net return for ease of interpretation – and on the product of the two $a_{T-1}r_{p,T}$; the parameter values are the same as those in the caption for Figure 2. We obtain

$$a_T = 0.2 + 0.63a_{T-1} - 0.35r_{p,T} + 0.75a_{T-1}r_{p,T} + \varepsilon_T, \tag{30}$$

with an R^2 of 43.7%. The relationship in (30) captures four features of the model-free system, which are most easily illustrated with numerical examples.

First, equation (30) shows that the model-free allocation at time T is closely tied to the previous period's allocation. For example, if $a_{T-1} = 20\%$ and the net portfolio return is a neutral $r_{p,T} = 0$, then the expected time T allocation $E(a_T) = 33\%$; and if $a_{T-1} = 80\%$ and $r_{p,T} = 0$, then $E(a_T) = 71\%$. In both cases, the expected time T allocation is fairly close to the time $T-1$ allocation. This is because, at each time, the model-free system primarily updates the Q value of the most recently-chosen action. As such, the Q values at time T are similar to the Q values at time $T-1$; this, in turn, means that the time T allocation is likely to resemble the time $T-1$ allocation.

Second, the numbers in the previous paragraph show that there is mean-reversion in the model-free allocation. Again, when the net portfolio return is $r_{p,T} = 0$, a prior allocation of 20% leads to an expected allocation of 33%, while a prior allocation of 80% leads to an expected allocation of 71%. The mean-reversion is due to the probabilistic action choice and to the fact that the set of possible allocations is bounded by 0% and 100%. If the investor has a high allocation to the stock market at time $T-1$, then, since there is a random component to his time T allocation and since this allocation must be between 0% and 100%, his time T allocation will on average be lower than at time $T-1$.

Third, regression (30) captures the reinforcing effect of the portfolio return. If $a_{T-1} = 20\%$ and the portfolio return is a neutral $r_{p,T} = 0$, then $E(a_T) = 33\%$; but if $r_{p,T} = 0.2$, then $E(a_T) = 29\%$: the high portfolio return reinforces the 20% allocation and pulls the time T allocation towards it, from 33% down to 29%. Similarly, if $a_{T-1} = 80\%$ and $r_{p,T} = 0$, then $E(a_T) = 71\%$; but if $r_{p,T} = 0.2$, then $E(a_T) = 76\%$: this time, the high portfolio return reinforces the 80% allocation and pulls the time T allocation towards it, from 71% up to 76%.

Finally, regression (30) captures the indirect way that the market return affects the model-free allocation. If $a_{T-1} = 20\%$ and the net *market* return is a neutral $r_{m,t} \equiv R_{m,t} - 1 = 0$, then $r_{p,T} = 0$ and $E(a_T) = 33\%$ as before. But if $r_{m,T} = 0.2$, then $r_{p,T} = 0.04$ and $E(a_T) = 32\%$. In this case, the high market return modestly reinforces the prior allocation of 20% and lowers the expected allocation by 1%, from 33% to 32%. Similarly, if $a_{T-1} = 80\%$ and $r_{m,T} = 0$, then $r_{p,T} = 0$ and $E(a_T) = 71\%$. But if $r_{m,T} = 0.2$, then $r_{p,T} = 0.16$ and $E(a_T) = 75\%$. In this case, the high market return strongly reinforces the prior allocation of 80% and increases the expected allocation by 4%, from 71% to 75%. Averaging across the two prior allocations, the positive market return increases the investor's time T allocation by approximately $(4-1)/2 = 1.5\%$. This provides a numerical illustration of the mechanism described in the previous section: the model-free allocation depends positively on past market returns because a high market return generates greater reinforcement when the investor's prior allocation is high.[26]

3.4 Variability and dispersion

Regression (30) in the previous section shows that model-free allocations are "sticky": the allocation at any time hews closely to the allocation in the previous period. This suggests that the model-free system will lead to less variability in an investor's allocation over time. We now document this more formally.

To demonstrate the result, we first allow for dispersion in learning rates across investors.[27] For each investor, we draw each of their learning rates – each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} – from a uniform distribution centered at $\bar{\alpha}$ and with width Δ. As before, the parameter values are $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$; and as in Section 3.1, we have $b = 0.0577$, so that there is some generalization. We set the new parameter Δ to 0.5. We take 1,000 investors, expose them to the same sequence of stock market returns from $t = -L$ to $t = T$, and compute the variability in their allocations: for each investor

[26] A more precise calculation, which averages across all eleven possible prior allocations, leads to a similar result. The 1.5% number approximately matches the prediction of the dashed line in Figure 2 for the sensitivity of the model-free allocation to a 20% net market return, namely $(0.1072)(0.2) = 2.14\%$, where 0.1072 is the coefficient on the most recent market return in a regression of model-free allocations on past market returns.

[27] Data on investor beliefs about future stock market returns suggest that there is substantial dispersion in learning rates across investors. Giglio et al. (2021) analyze such data and find that an individual fixed effect explains more of the variation in beliefs than a time fixed effect: some investors are persistently optimistic while others are persistently pessimistic. Capturing this in our framework requires substantial dispersion in learning rates across investors, a claim we have confirmed in simulated data: as we increase this dispersion, individual fixed effects explain more of the variation in beliefs. Intuitively, investors with high α_+^{MB} and low α_-^{MB} are persistently optimistic, while those with low α_+^{MB} and high α_-^{MB} are persistently pessimistic.

in turn, we compute the standard deviation of his allocations $\{a_{T-j}\}_{j=0}^{30}$ over time and then average these standard deviations across investors. We repeat this exercise 300 times for different return sequences and average the resulting variability measures.

The solid and dashed lines in the top three graphs in Figure 5 plot the variability of investor allocations under the model-based and model-free systems, respectively, as we vary the values of three parameters – the exploration parameter β, the mean learning rate $\bar{\alpha}$, and the dispersion Δ of learning rates – while keeping the other parameter values fixed at their benchmark levels. The graphs confirm that the model-free system leads to lower variability than the model-based system: the dashed lines are substantially below the solid lines.

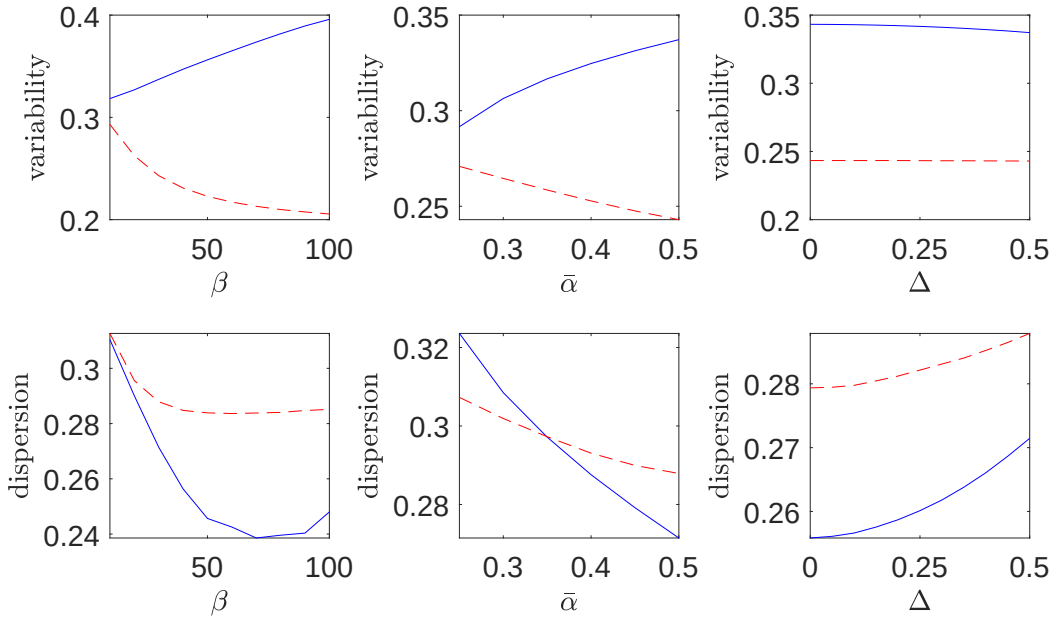

Figure 5. The upper graphs plot the variability of stock market allocations – the standard deviation of allocations between time 0 and time T, computed for each investor in turn and averaged across investors. The lower graphs plot the dispersion, across investors, of their stock market allocations at time T. The solid and dashed lines correspond to the model-based and model-free systems, respectively. For each system, the graphs vary the exploration parameter β, the mean learning rate $\bar{\alpha}$, or the dispersion in learning rates Δ, while keeping the other parameter values fixed at their benchmark levels. The results are averaged across 300 simulations; each simulation features 1,000 investors, all of whom see the same return sequence. The benchmark parameter values are $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

Using the simulated data from the above exercise, we also compute another quantity that will be helpful when we turn to applications, namely the dispersion in allocations across investors at time T. For each of our 300 simulations, and for each of the model-free and model-based systems, we compute the standard deviation of investors' time T allocations and then average these estimates across the 300 simulations. The lower graphs in Figure 5 plot the resulting dispersion measures as we vary each of β, $\bar{\alpha}$, and Δ. The graphs show that the two systems generate similarly high dispersion in investor allocations. We return to this finding in Section 4.

4 Applications

We now build on the analysis of Section 3 to show that our framework can shed light on a range of facts in finance. This is striking, for two reasons. First, in prior research, this framework has been used primarily to explain behavior in simple experimental settings; it is notable, then, that it can also account for real-world financial behavior. Second, one component of the framework is "model-free," and, as such, uses very little information about the nature of the task. It is striking that a framework that "knows" so little about financial markets can nonetheless help explain investor behavior in these markets.

We have associated the risky asset in our framework with the aggregate stock market. Our applications therefore focus on important facts about this market – facts about investor behavior, such as extrapolative demand and experience effects; facts about investor beliefs about market returns and the relationship between beliefs and allocations; and facts about asset prices, such as excess volatility. Through the model-based system, our framework preserves a role for beliefs in driving investor behavior and asset prices. However, through the model-free system, it introduces a new way of thinking about these facts, one based on reinforcement of past actions.

In what follows, we show that a simple parameterization of our framework can qualitatively, and even quantitatively, address a range of empirical facts. By "simple," we mean that, in this parameterization, each investor's learning rates α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} are constant over time; and, for all investors, the values of these learning rates are drawn from the same distribution. Our initial goal is not to provide a close quantitative fit to observed facts; it is to show that a simple parameterization can provide a qualitative, and approximate quantitative, fit to the data. Toward the end of this section, we estimate the parameter values that provide a closer quantitative match to the data.

To study the various applications, we start with the setup of Section 3. There are again $L+T+1$ dates, $t = -L,\ldots,-1,\ 0,\ 1,\ldots,\ T$. Relative to Section 3, we make one modification to make the framework more realistic: we allow for different cohorts of investors who enter financial markets at different times. Specifically, we take $L = T = 30$ and consider six cohorts, each of which contains $50,000$ investors, for a total of $300,000$ investors. The first cohort begins participating in financial markets at time $t = 0$; we track their allocation decisions until time $t = T$. For these investors, their model-based systems operate over the full timeline starting at time $t = -L$, but their model-free systems operate only from time $t = 0$ on. The second cohort enters at time $t = 5$; we track them until time $t = T$. For this cohort, the model-based system again operates over the full timeline starting at $t = -L$, but the model-free system operates only from time $t = 5$ on. The four remaining cohorts enter at dates $t = 10$, 15, 20, and 25.

Given the above structure, at time T, the cross-section of investors resembles the one we see in reality, namely one where investors differ in their number of years of participation in financial markets. As such, most of our analyses will focus on investor allocations at time T and on how these relate to other variables, such as investor beliefs at that time or the past stock market returns investors have been exposed to. For most of the applications, we conduct simulations in which each investor interacts with a different return sequence from time $t = -L$ to time $t = T$.

Each investor in the economy is trying to solve the problem in (10) and chooses allocations from the set $\{0\%,\ 10\%,\ldots,\ 90\%,\ 100\%\}$ according to the hybrid system in (24)-(25). For each investor, we draw the values of the learning rates α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} independently from a uniform distribution with mean $\bar{\alpha}$ and width Δ. We use the same parameter values throughout much of this section in order to show that a single parameterization is consistent with a range of empirical facts. As in Section 3, we set $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$, so that investors put equal weight on the model-free and model-based systems. Later, we formally estimate the value of w that best fits the data.

4.1 Extrapolative demand and excess volatility

Our first application builds on the analysis of Section 3.2. A common assumption in psychology-based models of asset prices and investor behavior is that people have extrapolative demand: their demand for a financial asset depends positively on the asset's past

returns, and especially on its recent past returns.[28]

The framework of Section 2 provides a new foundation for such extrapolative demand. As shown in Section 3.2, for a wide range of parameter values, the model-free and model-based systems both generate an allocation to the stock market that depends positively on past market returns and more so on recent past returns. To be clear, the mechanism in the case of the model-based system is similar to others that have been proposed. However, in the case of the model-free system, the mechanism is new to the finance literature. We explained the logic in full in Section 3.2; a brief summary is: Following a good stock market return, the reward prediction error is larger if the investor previously had a high allocation to the stock market than if he had a low allocation; a high allocation therefore receives more reinforcement, making it more likely that he will choose a high allocation going forward.

To confirm that the framework of Section 2 generates extrapolative demand, we run a regression of investors' allocations a_T at time T, as determined by the hybrid system, on the past stock market returns each of them has observed. The relationship between the allocation and past returns is plotted as the solid line in Figure 6. The graph confirms that an investor's allocation to the stock market is a positive function of the market's past returns, with weights that decline the further back we go into the past.

The solid line in Figure 6 is similar to the line marked "Hybrid" in Figure 2 in that both lines correspond to decisions made under the hybrid system. However, the two lines differ because, relative to the analysis in Section 3.2, we are now allowing for dispersion across investors in their learning rates and for multiple cohorts. The multiple cohorts in particular make the solid line in Figure 6 decline more quickly than the "Hybrid" line in Figure 2: some of the investors in the market at time $T = 30$ entered only at time 25; as such, their model-free system puts no weight on returns before time 25.

The framework of Section 2 offers another insight relative to existing finance research on extrapolative demand, namely that this demand has two sources which operate on different time scales: a model-based source that puts heavy weight on *recent* returns, and a model-free source that puts substantial weight even on *distant* past returns. We make use of this two-component structure of extrapolative demand in subsequent applications.

[28]A partial list of papers that study extrapolative demand, either theoretically or empirically, is Cutler, Poterba, and Summers (1990), De Long et al. (1990), Barberis and Shleifer (2003), Barberis et al. (2015, 2018), Cassella and Gulen (2018), Chen, Liang, and Shi (2022), Jin and Sui (2022), Liao, Peng, and Zhu (2022), and Pan, Su, and Yu (2022).

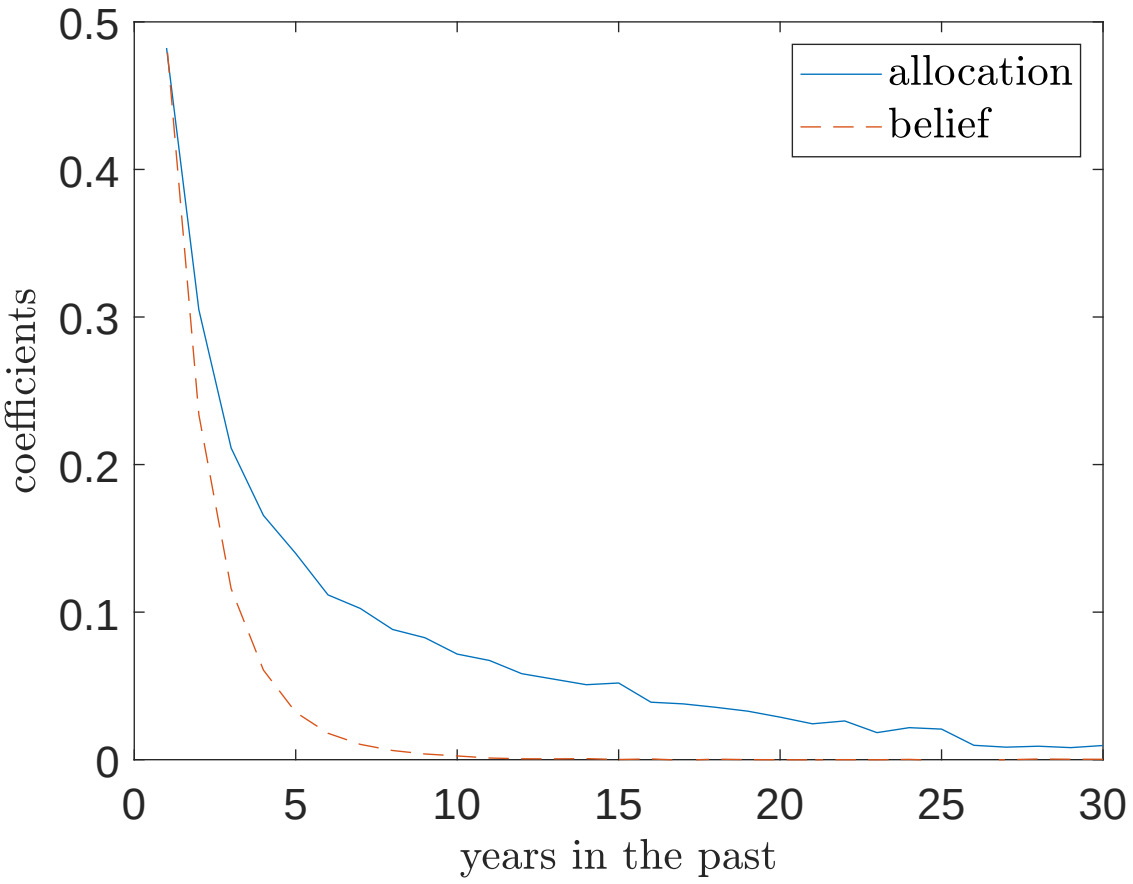

Figure 6. The solid line plots the coefficients in a regression of the stock market allocation a_T at date T chosen by investors who use a hybrid system to make decisions on the past 30 years of stock market returns the investors were exposed to. The dashed line plots the coefficients in a regression of investors' expectations at time T about the future one-year stock market return on the past 30 years of stock market returns. There are 300,000 investors: six cohorts of 50,000 investors each who enter financial markets at different times. For each investor, each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} is drawn independently from a uniform distribution with mean $\bar{\alpha}$ and width Δ. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$.

Prior work has shown that certain types of extrapolative demand can help explain the excess volatility in stock markets and the predictability of market returns (De Long et al., 1990; Barberis et al., 2015). Given that the investors in our framework have a form of extrapolative demand, it is natural to expect that, in an asset pricing setting, these investors will generate excess volatility and predictability. Our focus in this paper is on investor behavior, not asset prices. Nonetheless, in Online Appendix D, we present a simple model of asset prices in which some investors have the two-component structure of extrapolative

demand generated by model-free and model-based learning, and confirm that we observe excess volatility and return predictability. Our framework preserves a role for beliefs in generating excess volatility – this comes from the model-based system – but also points to a new mechanism that may contribute to this volatility, namely model-free reinforcement of past actions.

4.2 Experience effects

Malmendier and Nagel (2011) show that investors' decisions are affected by their experience: whether an investor participates in the stock market, and how much he allocates to the stock market if he does participate, can be explained in part by the stock market returns he has personally experienced – in particular, by a weighted average of the returns he has personally lived through, with more weight on more recent returns.

The framework of Section 2 provides a foundation for such experience effects. Since the model-free system engages only when an investor is actively experiencing financial markets, the framework predicts that investors who enter financial markets at different times, and who therefore experience different returns, will choose different allocations.

There are two key features of experience effects that we aim to capture. The first is that, if an investor begins participating in financial markets at time t, his subsequent allocations to the stock market should depend substantially more on the stock market return at time $t+1$, $R_{m,t+1}$ – a return he experienced – than on the stock market return at time t, $R_{m,t}$, a return he did not experience. Put differently, if we plot the coefficients in a regression of investor allocations on past market returns, we should see a "kink" in the coefficients at the moment the investor enters financial markets. The second feature of experience effects is that the coefficients in a regression of investor allocations on past experienced stock market returns should decline for more distant past returns. To capture both features, Malmendier and Nagel (2011) propose that investors' decisions are based on a weighted average of past returns in which, for an investor at time t with n years of experience, the weight on the return j years earlier, $R_{m,t+1-j}$, is

$$(n+1-j)^{\lambda}/A, \qquad j=1,2,\ldots,n, \tag{31}$$

where λ is estimated to be approximately 1.3 and A is a normalizing constant, and where the weight on returns the investor did not experience is zero.

To see if our framework can generate these two features of experience effects, we proceed

as follows. For each of the six cohorts, we take the 50,000 investors in the cohort and regress their time T allocations a_T on the past 30 years of stock market returns. Figure 7 presents the results. The six graphs correspond to the six cohorts. In each graph, the solid line plots the coefficients in the above regression, normalized to sum to one so that we can compare them to the Malmendier and Nagel (2011) coefficients in (31). The dashed line plots the functional form in (31) for the cohort in question, and the vertical dotted line marks the point at which the cohort enters financial markets.

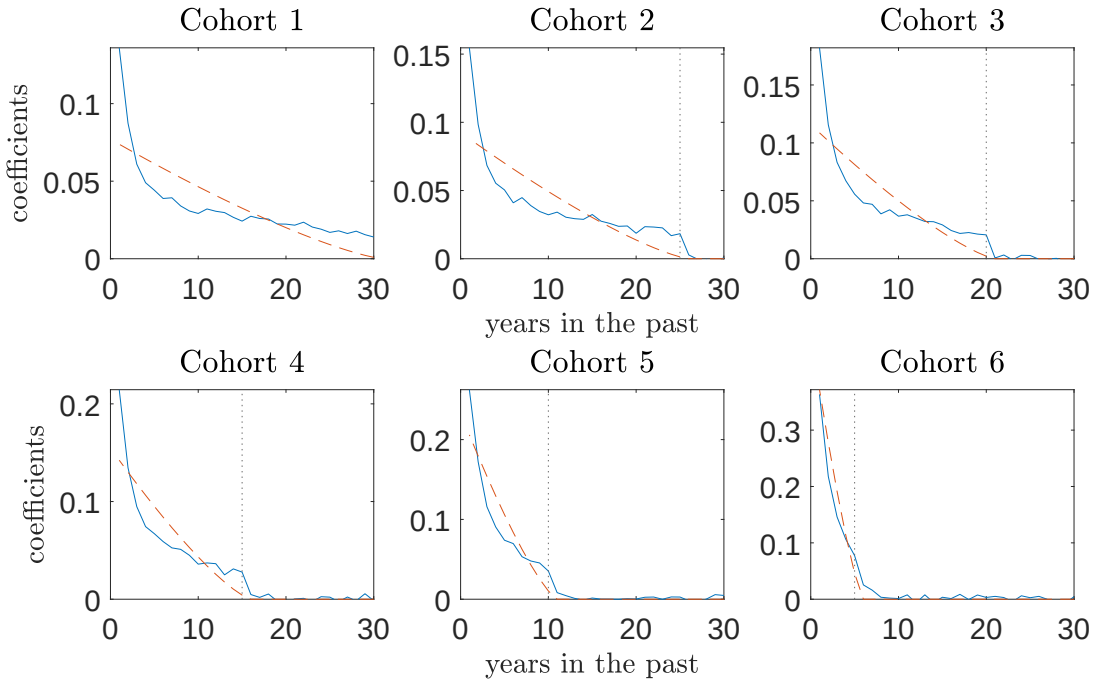

Figure 7. The six graphs correspond to six cohorts of investors. In each graph, the solid line plots the coefficients, normalized to sum to one, in a regression of the time T stock market allocations a_T of the investors in that cohort on the past 30 years of stock market returns they were exposed to. The six cohorts have different numbers of years of experience, namely $n = 5$, 10, 15, 20, 25, and 30; the vertical dotted line in each graph marks the time at which the cohort enters financial markets. There are 300,000 investors, with 50,000 in each cohort. For each investor, each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} is drawn independently from a uniform distribution with mean $\bar{\alpha}$ and width Δ. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$. In each graph, the dashed line plots a functional form for experience effects calibrated to data by Malmendier and Nagel (2011), namely $(n + 1 - j)^\lambda / A$, where j is the number of years in the past, $\lambda = 1.3$, and A is a normalizing constant.

By comparing the solid and dashed lines for each graph in turn, we see that our framework can capture both aspects of experience effects. Consider the bottom-left graph for cohort 4 which enters at date 15. The solid line shows that our framework generates a kink in the dependence of allocation on past market returns as we move from a return these investors experienced – the return 15 years in the past – to one they did not experience, the return 16 years in the past. The kink is driven by investors' model-free system, which puts substantial weight even on a return experienced 15 years in the past, but no weight at all on returns before that. The graph also shows that, within the subset of returns that these investors experience, their allocation puts greater weight on more recent past returns. Both the model-free and model-based systems contribute to this pattern, although the model-based system does so more.

Similar patterns can be seen in the other graphs. In each case, the solid line exhibits a kink at the moment that the investors in that cohort begin experiencing returns; and within the subset of returns that the investors in that cohort experience, there is more weight on more recent returns.

We have presented a foundation for experience effects based on model-free and model-based learning. It is instructive to compare this to an alternative foundation rooted in memory (Bordalo, Gennaioli, and Shleifer, 2020; Wachter and Kahana, 2022; Malmendier and Wachter, 2022). Specifically, in Wachter and Kahana (2022), an experienced return is treated differently from a return that is not experienced because it is processed by a separate memory system. In our framework, an experienced return is treated differently because it engages a distinct learning algorithm, model-free learning. In Wachter and Kahana (2022), the investor's allocation can put substantial weight even on a return in the distant past. This is because the "context" that is associated with observed data at each moment is assumed to change slowly over time. In our framework, it is instead because the model-free system learns slowly, at each time primarily updating only the value of the most recently-chosen action.

Wachter and Kahana (2022) also obtain a "jump back in time" effect whereby today's data remind the investor of an earlier context in which the same data were observed, thereby recruiting additional data from that earlier context. Once we allow for state dependence, our framework generates a similar effect: if the investor encounters state s_t at time t, his action is influenced by the reward he experienced at earlier times when he was again in state s_t. We analyze a state-dependent version of our framework in the Online Appendix.

4.3 Investor beliefs: Overreaction and the frequency disconnect

Individual investors overreact to recent market returns when forming beliefs about future market returns: their beliefs are a positive function of recent returns even though there is little autocorrelation in realized returns (Greenwood and Shleifer, 2014). Our framework captures this overreaction. As we confirm below, after a high return, the model-based system increases the probability it assigns to good returns, leading the investor to expect a higher return in the future. In the same way, the framework can capture the overreaction we observe more generally for low-persistence processes (Bordalo et al., 2020; Afrouzi et al., 2023). For example, if the model-based system is trying to build a probability distribution for earnings growth, then, following high earnings growth, it raises the probability it assigns to good earnings growth outcomes, and thus over-estimates future earnings growth.

Aside from capturing overreaction in beliefs, our framework also resolves two puzzling disconnects between investor beliefs and investor actions – one in the frequency domain, which we discuss in this section, and one in the cross-section of investors, which we address in the next section. We account for these puzzles by way of a deep property of our framework, namely that, of the two systems, only the model-based system has an explicit role for beliefs. The model-free system, by contrast, has no notion of beliefs: it does not construct a probability distribution of future outcomes; instead, it learns the value of actions simply by trying them and observing the outcomes.

The disconnect in the frequency domain is simple to state. While investor expectations about future returns depend heavily on *recent* past returns, investor stock market allocations depend to a substantial extent even on distant past returns (Malmendier and Nagel, 2011).[29]

Two features of our framework allow it to explain this disconnect. First, as noted above, only the model-based system has an explicit role for beliefs. Second, relative to the model-based system, the model-free system recommends allocations that put substantially more weight on distant past returns. Taken together, these features mean that, when an investor is asked for his beliefs about future stock market returns, he necessarily consults the model-based system – the only system that can answer the question – and therefore gives a response that puts heavy weight on recent returns. By contrast, his allocation is based on both systems and therefore puts greater relative weight on distant past returns. As such, the framework

[29]We can formalize this in the following way. When Malmendier and Nagel (2011) use the weights in (31) to characterize the relationship between an investor's allocation and the past returns he has experienced, they obtain an estimate of $\lambda \approx 1.3$. Suppose that we now take the functional form in (31) and use it, with $n = 30$, to characterize the relationship between investor *beliefs* and the past 30 years of stock market returns. Using Gallup data on stock market expectations from October 1996 to November 2011, we find that the best fit is for $\lambda \approx 37$, which puts much more weight on recent returns.

drives a wedge between actions and beliefs.

Figure 6 illustrates these points. As discussed in Section 4.1, the solid line shows how *allocations* depend on past returns: it plots the coefficients in a regression of investors' allocations to the stock market at time T on the past 30 years of stock market returns they were exposed to. The dashed line shows how *beliefs* depend on past returns: it plots the coefficients in a regression of investors' expectations at time T about the future one-year stock market return on the past 30 years of stock market returns they were exposed to. Comparing the two lines, we see that, while beliefs depend primarily on recent returns, allocations depend significantly even on distant past returns.

A number of studies find a positive time-series correlation between investor beliefs and allocations. For example, Greenwood and Shleifer (2014) find that the average investor expectation of future stock market returns is positively correlated with net flows into equity-oriented mutual funds. Our framework is consistent with such findings: in our simulated data, there *is* a positive time-series correlation between investor allocations and beliefs, both at the individual and aggregate levels. However, underlying the positive correlation in actual data is a frequency disconnect, with beliefs putting more weight on recent returns than do allocations; it is this puzzling disconnect that our framework can explain.

4.4 Investor beliefs: The cross-sectional disconnect

Using survey responses from Vanguard investors, as well as data on these investors' allocations to the stock market, Giglio et al. (2021) document another disconnect between beliefs and actions. Regressing investors' stock market allocations on investors' expected one-year stock market returns, they obtain a coefficient approximately equal to one. By contrast, a traditional Merton model of portfolio choice predicts a much higher coefficient.

Our framework can help explain this disconnect. The mechanism is similar to that for the frequency disconnect: it again relies on the fact that, while an investor's allocation is based on both the model-free and model-based systems, only the model-based system has an explicit role for beliefs. To see the implications of this, suppose that the stock market posts a high return. The investor's expectation about the future stock market return will then go up significantly: the model-based system, which determines beliefs, puts substantial weight on recent returns. However, the investor's allocation will be less sensitive to the recent return: it is determined in part by the model-free system, which, relative to the model-based system, puts much less weight on recent returns.

We can examine this effect quantitatively. In simulated data, we run a regression of investors' stock market allocations at time T on their expected returns on the stock market over the next year. For our benchmark parameter values, and specifically for our benchmark value of $w = 0.5$, the regression coefficient is 1.12; this is similar to that obtained by Giglio et al. (2021) in actual data and confirms that our framework can help explain the cross-sectional disconnect. The model-free system plays an important role in this result: if we increase the weight on the model-free system from 0.5 to 0.9, say, the sensitivity of allocations to beliefs falls even further, from 1.12 to 0.45. The probabilistic choice plays a minor role in this result: even if we turn it off, the framework continues to generate an allocation-belief sensitivity that is low, and all the lower, the greater the weight on the model-free system.

The low sensitivity of allocations to beliefs initially appears puzzling in light of the excess volatility in the stock market. However, our framework shows that they can co-exist: as outlined above, the investors in our framework exhibit a low allocation-belief sensitivity; but because their demand has an extrapolative structure, it nonetheless generates excess volatility in asset prices, as shown in Online Appendix D.

In Section 2.5, we noted that, according to a well-known hypothesis in psychology, the brain puts more weight on the learning system it views as more reliable. This idea – one that we analyze formally in Online Appendix A – leads to one of the clearest predictions of our framework, namely that, if an individual is more confident in his beliefs, his actions will be more sensitive to his beliefs. Confidence in one's beliefs is a sign that the model-based system – the system that generates beliefs – is more reliable. The brain therefore allocates more control to it, leading to a higher sensitivity of allocations to beliefs. Giglio et al. (2021) offer some evidence that is consistent with this: they find that, for the subsample of people who say they are more confident in their beliefs, allocations are indeed more sensitive to beliefs.

4.5 Dispersion, inertia, and inelasticity

Households differ in their asset allocations – in particular, the fraction of wealth invested in the stock market varies substantially from one household to another. Economists typically attribute these differing allocations to differences in beliefs – differences in perceived expected returns or risk – or to differences in objective functions.

The lower panel of Figure 5 shows that the model-based system generates substantial dispersion in allocations. This dispersion is primarily due to differences in beliefs across

investors, which in turn are driven by differences in learning rates; the probabilistic choice further adds to the dispersion. The lower panel shows that the model-free system also generates substantial dispersion in allocations. This is striking because this dispersion cannot be easily attributed to differences in beliefs or objective functions: as noted earlier, the model-free system has no notion of beliefs; moreover, in our setting, all investors have the same objective function in (10). Instead, the differences in allocations recommended by the model-free system are due to the process of decision-making itself. The probabilistic choice leads investors to try different allocations in their early years of financial market participation. Different allocations are then reinforced for different investors, which leads to differences in allocations even many years later.

While there is substantial dispersion in households' actual allocations to the stock market, there is also individual-level inertia in these allocations over time (Agnew, Balduzzi, and Sunden, 2003; Ameriks and Zeldes, 2004). This inertia is often attributed to transaction costs, procrastination, or inattention.

The framework in this paper offers a new way of thinking about inertia in investor holdings: it says that the inertia arises endogenously from the model-free system. Regression (30) in Section 3.3 shows that an investor's model-free allocation in any period is closely tied to his allocation in the previous period. More directly, the upper panel of Figure 5 shows that, relative to the model-based system, the model-free system generates lower variability, or equivalently, higher inertia. The reason is that the model-free system learns slowly: at each time, it primarily updates only the value of the most recently-chosen allocation. This, in turn, increases the likelihood that the allocation at time t will be similar to the allocation at time $t-1$.

The inertia generated by the model-free system also offers a foundation for the market inelasticity documented by Gabaix and Koijen (2022) – the finding that, if some investors have uninformed demand for an asset that pushes up its price, other investors do not absorb the demand to the extent predicted by traditional models. Gabaix and Koijen (2022) note that one possible source of inelasticity is investment mandates that constrain the holdings of asset managers. Our framework points to an alternative source. Since the model-free system learns slowly, it generates inertia in investors' allocations, which in turn reduces the extent to which they will respond to an uninformed demand shock.

4.6 Non-participation

For our final two applications – non-participation and persistent investment mistakes – we use modified versions of our framework that better fit the context at hand.

A long-standing question asks why many U.S. households do not participate in the stock market despite its substantial risk premium. Our framework can shed light on this. In particular, the model-free system tilts investors toward not participating. To see why, consider an investor who makes decisions according to the model-free system. If he allocates some money to the stock market but then experiences a poor market return, this lowers the Q values of the chosen allocation and of those similar to it; this, in turn, raises the probability that, in a subsequent period, he will switch to a 0% allocation to the market. Importantly, if he does move to a 0% allocation, the model-free system will update only the Q value of the 0% allocation: generalization aside, it learns only about the action taken. As such, it stops learning about the stock market and, in particular, fails to learn that the stock market has better properties than indicated by the one poor return the investor experienced. This will tend to keep the investor at a 0% allocation for an extended period of time.

We illustrate this in a modified version of our framework with just two allocations: 0% and 100%. It is natural to use a two-allocation framework for this application because the participation decision has a binary flavor: Should I participate or not? In addition, because the multi-cohort structure we used earlier does not play an interesting role in this application, we consider a single cohort of investors who enter financial markets at time 0.

We take $300,000$ investors and expose each of them to a different sequence of stock market returns. For each investor, we compute the fraction of time between dates 1 and T that he chooses a 0% allocation. In addition, for each investor, we identify the episodes where he allocates 0% to the stock market for multiple consecutive years and record the duration of the longest such episode. We do this exercise twice: first for the case where decisions are made by the model-free system and then for the case where they are made by the model-based system. The parameter values are the same as before, namely $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

The results confirm that the model-free system tilts investors toward non-participation. We find that, under the model-free system, 43% of investors are not participating – in other words, are at a 0% allocation – for at least 80% of the 30 dates from $t = 1$ to $t = 30$. By contrast, under the model-based system, fewer than 1% of investors spend more than 80% of the time not participating. In a similar vein, under the model-free system, 59% of investors have a non-participation streak that is at least 10 years long; under the model-based system,

only 7% of investors have a streak of this length. The simulated data also support the mechanism for non-participation laid out above. We find that, under the model-free system, long streaks of non-participation are typically preceded by a poor experienced stock market return. The longer the non-participation streak, the more negative the prior experienced return, on average.

4.7 Persistent investment mistakes

Many households make suboptimal financial choices; moreover, they often persist in these poor choices for many years. Our framework can help explain this. The idea is simple: The model-free system learns slowly; at each date, it learns primarily about the value of the action a person is currently taking. As a result, it can take a long time to learn the optimal course of action.

To demonstrate this quantitatively, it is natural to consider a slightly different setting from the one we have used so far. In this new setting, there are ten risky assets and no risk-free asset. The gross return on asset i from time $t-1$ to t, $R_{i,t}$, is distributed as

$$\log R_{i,t} \sim N(\mu_i, \sigma_i^2), \text{ i.i.d. over time,}$$

and the returns on the ten assets are uncorrelated with each other. For all ten assets, $\sigma_i = 0.2$, but while assets 1 through 9 have the same low $\mu_i = 0.01$, asset 10 has a substantially higher $\mu_{10} = 0.06$. Analogous to the goal function in (10), each investor's objective is to maximize the expected sum of discounted log portfolio returns where, at each time, he can invest his wealth in just one of the ten risky assets. The question is: At time $T = 30$, what fraction of investors recognize that asset 10 is the best option, in the sense that they assign it the highest estimated Q value among the 10 options?

We answer this question separately for the model-based system and for the model-free system. In this section, we adjust the model-based system so that it serves as a normative benchmark by which to judge the efficiency of the model-free system. We endow the model-based system with some knowledge of the environment, namely that each asset's return is i.i.d. over time and uncorrelated with the returns of other assets – but no other information beyond this. All investors use declining model-based learning rates equal to those in (26); for these learning rates, consistent with the i.i.d. assumption, investors are equally weighting the past returns on each asset.

The model-free system operates as before: all investors use constant learning rates; for

each investor, his learning rates are drawn from a uniform distribution with mean $\bar{\alpha} = 0.5$ and width $\Delta = 0.5$, and the exploration parameter is $\beta = 30$. The remaining parameters are the same for both systems: there are $300,000$ investors in each case; and $L = 0$, $T = 30$, $\gamma = 0.97$, and $b = 0$, so that there is no generalization. Setting $L = 0$ means that there are no data prior to time 0 that the model-based system can learn from; this puts the two systems on more equal footing. In Online Appendix E, we present the full updating equations for both systems.

We find that, in the case of the model-based system, at time $T = 30$, 47% of investors recognize that asset 10 is the best option: they assign it the highest Q value among the 10 options. By contrast, for the model-free system, at time $T = 30$, just 20% of investors recognize that asset 10 is the best option. Consistent with our claim above, then, the model-free system learns slowly: since, at each time, it updates only the value of the most recently-chosen action, it takes much longer to figure out the sensible course of action.

While our analysis is based on a setting with ten risky assets, we expect the findings of this section to apply more generally to any situation where an investor faces a number of possible courses of action and has to figure out which one is best. Since the model-free system learns slowly, it takes the investor a long time to discover the best option; even after many years, he may still be investing suboptimally.[30]

4.8 Parameter estimation

Throughout Section 4, we have taken a simple parameterization of our framework and shown that it provides a qualitative and approximate quantitative match to a number of facts about investor behavior. We now estimate the parameter values that best match the data. We have three empirical targets: the relationship between past returns and investor beliefs about future returns, as measured from surveys of investors; the sensitivity of allocations to beliefs, as computed by Giglio et al. (2021); and the dependence of allocations on past returns, as reported by Malmendier and Nagel (2011) in their analysis of experience effects. The parameters we estimate are the mean model-based learning rate across investors $\bar{\alpha}^{MB}$; the mean model-free learning rate $\bar{\alpha}^{MF}$; the exploration parameter β; and most important, the weight w on the model-based system.

We explain the estimation procedure in full in Online Appendix F and summarize it here. We do the estimation in two steps. We first use data on investor beliefs to estimate

[30]See Gagnon-Bartsch, Rabin, and Schwartzstein (2021) for an alternative approach to slow learning, one based on inattention to useful information.

$\bar{\alpha}^{MB}$. We then estimate $\bar{\alpha}^{MF}$, β, and w by targeting the allocation-belief sensitivity and the experience effect. We keep the remaining parameters at their benchmark values from before, namely $L = T = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.[31]

In the first step, we estimate the mean model-based learning rate $\bar{\alpha}^{MB}$ by searching for the value of this parameter that best fits the empirical relationship between investor beliefs and past market returns. Specifically, we take monthly Gallup data from October 1996 to November 2011 on average investor beliefs about the future one-year stock market return and regress these beliefs on past annual stock market returns. We search for a value of $\bar{\alpha}_{MB}$ that, in simulated data from the model-based system, best matches the regression coefficients from the Gallup data. We find this to be $\bar{\alpha}^{MB} = 0.33$.

With this value of $\bar{\alpha}^{MB}$ in hand, we move to the second step: we search for values of $\bar{\alpha}^{MF}$, β, and w that best match two empirical targets. The first is the coefficient in a regression of investor allocations on investor beliefs, which Giglio et al. (2021) find to be approximately one. For given values of $\bar{\alpha}^{MF}$, β, and w, we can compute this coefficient in simulated data from our framework. Our second target is the functional form in (31) with $\lambda = 1.3$, which Malmendier and Nagel (2011) use to capture the relationship between allocations and past returns. Intuitively, we are looking for values of $\bar{\alpha}^{MF}$, β, and w that minimize the distance between unnormalized versions of the solid and the dashed lines in the six graphs in Figure 7.

We find that the parameter values that best match the allocation-belief sensitivity and the experience effect are $\bar{\alpha}^{MF} = 0.26$, $\beta = 20$, and $w = 0.38$. In words, to match the data, our framework requires substantial weight on both the model-free and model-based systems. The reason is the following. As shown by the dashed lines in Figure 7, the experience effect we are trying to capture involves both a steep initial decline in the coefficients on past returns, but also a significant dependence on distant past experienced returns. The upper panel of Figure 3 shows that the model-based system can capture the steep initial decline in coefficients, but, when calibrated to do so, it cannot capture the dependence on distant past returns. By contrast, the lower panel of Figure 3 shows that the model-free system can capture a high dependence on distant past returns but not the initial decline. To match both parts of the experience effect, we need to put substantial weight on both systems. The significant weight on the model-free system also helps to explain the low sensitivity of allocations to beliefs.

[31] We have repeated the estimation analysis for other values of these parameters and find that our main result – that the data is best explained by a framework that puts substantial weight on both the model-free and model-based systems – continues to hold.

5 Some Extensions

We now discuss some extensions of our framework. We start with three extensions that we have studied in detail; then, more briefly, we comment on three extensions that we leave to future work.

Time-varying weights on the two systems. We have focused on the case where w, the weight on the model-based system, is constant over time. A well-known hypothesis in psychology is that w varies over time: at each moment, the brain puts more weight on the system that it deems more reliable (Daw, Niv, and Dayan, 2005). In Online Appendix A, we formalize this idea using an approach of Lee, Shimojo, and O'Doherty (2014) in which the reliability of a system is measured by the absolute magnitude of its past prediction errors, and explore its implications. We first show that, in our setting, an investor will gradually put more weight on the model-free system over time: as the system gains experience, it becomes more reliable and the brain assigns more control to it. This implies, for example, that the allocations of older investors will be less sensitive to recent market returns – this is consistent with the evidence in Malmendier and Nagel (2011) – and also less sensitive to their beliefs. The framework also predicts that, following a large absolute stock market return, the weight on the model-free system will go down: an extreme stock market return generates a large absolute reward prediction error which lowers the model-free system's perceived reliability.

Other model-free and model-based systems. The properties of the model-free Q-learning system we have documented in this paper are likely to be robust to using alternative model-free frameworks. The reason is that all model-free systems are similar at their core: the individual takes an action, and based on the outcome, he updates the value of the action. Consistent with this claim, in Online Appendix G, we replace Q-learning with SARSA, an alternative model-free framework, and show that it leads to similar predictions.

When specifying the model-*based* part of our framework, we have a much wider range of choices. In Section 2, we adopted a model-based system inspired by those used in psychology, but others are of course possible. For example, some investors may use a model-based system with a more contrarian flavor – one that, following a good stock market return, recommends a lower allocation to the stock market on the grounds that it may now be overvalued. Such a model-based system would create a new tension with the model-free system: after a good stock market return, the model-free system will want to increase exposure to the stock market while the model-based system will want to reduce it.

State dependence. Thus far, we have not allowed for state dependence: we consider

action values $Q(a)$ rather than state-action values $Q(s, a)$ because even this simple case has many applications. In Online Appendix H, we examine the predictions of our framework when we allow for state dependence – in particular, when there are two observable states and the mean stock market return differs across them. We find that the framework continues to exhibit the property that underlies a number of the applications in Section 4, namely that, relative to the model-based system, the model-free system puts significantly more weight on distant past returns.

There are three more extensions whose detailed analysis we leave to future work:

Time-varying learning rates. We have taken each investor's learning rates to be constant over time and have shown that even this simple case has many applications. Nonetheless, learning rates may vary over time. For example, there is evidence that they go up at times of greater volatility (Behrens et al., 2007). Such an assumption can be incorporated into our framework and may lead to useful new predictions – for example, about investor behavior during crisis periods.

Alternative action spaces. Throughout the paper, we have used a standard action space based on the fraction of wealth allocated to the stock market: at each time, an investor can allocate 0% of his wealth to the stock market, or 10%, or 20%, and so on. One feature of the model-free system is that it can easily accommodate alternative action spaces – for example, one with the three possible actions: "do nothing," "increase exposure to the stock market by 10%," and "decrease exposure to the stock market by 10%." We have incorporated this alternative action space into our framework and find that its implications are broadly similar to those we have outlined in the paper. We leave a fuller analysis of this topic to future work.

Inferring beliefs from the model-free system. Until now, we have associated beliefs only with the model-based system. However, it is possible that an individual may also use the model-*free* system to make inferences about beliefs. For example, when an investor is asked for his beliefs about the stock market's future return or risk, it is natural that he will consult the model-based system, which will give him a direct measure of beliefs. However, he may also be influenced by the model-free system, and if $Q^{MF}(a = 1) > Q^{MF}(a = 0)$, so that his model-free system assigns the stock market a higher Q value than the risk-free asset, he may take this as a sign that the stock market has better *properties*, on several dimensions – for example, both a higher expected return and lower risk. This can help explain Giglio et al.'s (2021) finding that, when investors expect high returns in the stock market, they simultaneously expect the market to have lower risk, contrary to the prediction of traditional frameworks in which subjectively perceived risk and return are positively related.

6 Conclusion

When economists try to explain human decision-making in dynamic settings, they typically assume that people are acting "as if" they have solved a dynamic programming problem. By contrast, psychologists and neuroscientists are increasingly embracing a different approach, one based on model-free and model-based learning. In this paper, we import this framework into a simple financial setting, study its properties, and use it to account for a range of facts: facts about investor behavior, such as extrapolative demand and experience effects; facts about beliefs, such as overreaction in beliefs and the relationship between beliefs and stock market allocations; and facts about asset prices, such as excess volatility and return predictability. Through the model-based system, our framework preserves a role for beliefs in driving investor behavior and asset prices. However, through the model-free system, it also introduces a new way of thinking about these facts, one based on reinforcement of past actions.

The vast majority of economic frameworks take a model-based approach. Model-free reinforcement learning, by contrast, has had a much smaller footprint in economics and finance. The results in this paper challenge this state of affairs: they suggest that model-free learning may be more common in economic settings than previously realized.

There are two broad directions for future research. We can apply the framework proposed here to other economic domains. We can also incorporate richer psychological assumptions – for example, about time-varying learning rates, time-varying weights on the two systems, or state dependence. We expect that both of these broad directions will prove fruitful and will shed new light on people's choices in economic settings.

7 References

Afrouzi, H., Kwon, S., Landier, A., Ma, Y., and D. Thesmar (2023), "Overreaction in Expectations: Theory and Evidence," *Quarterly Journal of Economics*, forthcoming.

Agnew, J., Balduzzi, P., and A. Sunden (2003), "Portfolio Choice and Trading in a Large 401(k) Plan," *American Economic Review* 93, 193-215.

Allos-Ferrer, C. and M. Garagnani (2022), "Part-time Bayesians: Incentives and Behavioral Heterogeneity in Belief Updating," *Management Science*, forthcoming.

Ameriks, J. and S. Zeldes (2004), "How Do Portfolio Shares Vary with Age?," Working paper.

Balleine, B., Daw, N., and J.P. O'Doherty (2009), "Multiple Forms of Value Learning and the Function of Dopamine," in *Neuroeconomics*, Academic Press.

Barberis, N., Greenwood, R., Jin, L., and A. Shleifer (2015), "X-CAPM: An Extrapolative Capital Asset Pricing Model," *Journal of Financial Economics* 115, 1-24.

Barberis, N., Greenwood, R., Jin, L., and A. Shleifer (2018), "Extrapolation and Bubbles," *Journal of Financial Economics* 129, 203-227.

Barberis, N. and A. Shleifer (2003), "Style Investing," *Journal of Financial Economics* 68, 161-199.

Behrens, T., Woolrich, M., Walton, M., and M. Rushworth (2007), "Learning the Value of Information in an Uncertain World," *Nature Neuroscience* 10, 1214-1221.

Benartzi, S. and R. Thaler (1995), "Myopic Loss Aversion and the Equity Premium Puzzle," *Quarterly Journal of Economics* 110, 73-92.

Bordalo, P., Gennaioli, N., and A. Shleifer (2020), "Memory, Attention, and Choice," *Quarterly Journal of Economics* 135, 1399-1442.

Bordalo, P., Gennaioli, N., Ma, Y., and A. Shleifer (2020), "Overreaction in Macroeconomic Expectations," *American Economic Review* 110, 2748-2782.

Camerer, C. (2003), *Behavioral Game Theory*, Russell Sage Foundation and Princeton University Press, Princeton, New Jersey.

Camerer, C. and T. Ho (1999), "Experience-weighted Attraction Learning in Normal-form Games," *Econometrica* 67, 827-874.

Cassella, S. and H. Gulen (2018), "Extrapolation Bias and the Predictability of Stock Returns by Price-scaled Variables," *Review of Financial Studies* 31, 4345-4397.

Charness and Levin (2005), "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect," *American Economic Review* 95, 1300-1309.

Charpentier, C. and J.P. O'Doherty (2018), "The Application of Computational Models to Social Neuroscience: Promises and Pitfalls," *Social Neuroscience* 13, 637-647.

Chen, W., Liang, S., and D. Shi (2022), "Who Chases Returns? Evidence from the Chinese Stock Market," Working paper.

Collins, A. (2018), "Learning Structures Through Reinforcement," in *Goal-directed Decision-making: Computations and Neural Circuits*, Academic Press.

Cutler, D., Poterba, J., and L. Summers (1990), "Speculative Dynamics and the Role of Feedback Traders," *American Economic Review Papers and Proceedings* 80, 63-68.

Daw, N. (2014), "Advanced Reinforcement Learning," in *Neuroeconomics*, Academic Press.

Daw, N., Gershman, S., Seymour, B., Dayan, P., and R. Dolan (2011), "Model-based Influences on Humans' Choices and Striatal Prediction Errors," *Neuron* 69, 1204-1215.

Daw, N., Niv, Y., and P. Dayan (2005), "Uncertainty-based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control," *Nature Neuroscience* 8, 1704-1711.

De Long, J.B., Shleifer, A., Summers, L., and R. Waldmann (1990), "Positive Feedback Investment Strategies and Destabilizing Rational Speculation," *Journal of Finance* 45, 375-395.

Dunne, S., D'Souza, A., and J.P. O'Doherty (2016), "The Involvement of Model-based but not Model-Free Learning Signals During Observational Reward Learning in the Absence of Choice," *Journal of Neurophysiology* 115, 3195-3203.

Erev, I. and A. Roth (1998), "Predicting How People Play in Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," *American Economic Review* 88, 848-881.

Evans, G. and S. Honkapohja (2012), *Learning and Expectations in Macroeconomics*, Princeton University Press, Princeton, New Jersey.

Frydman, C. and L. Jin (2022), "Efficient Coding and Risky Choice," *Quarterly Journal of Economics* 137, 161-213.

Gabaix, X. and R. Koijen (2022), "In Search of the Origin of Financial Fluctuations: The Inelastic Markets Hypothesis," Working paper.

Gagnon-Bartsch, T., Rabin, M., and J. Schwarzstein (2021), "Channeled Attention and

Stable Errors," Working paper.

Giglio, S., Maggiori, M., Stroebel, J., and S. Utkus (2021), "Five Facts about Beliefs and Portfolios," *American Economic Review* 111, 1481-1522.

Glascher, J., Daw, N., Dayan, P., and J.P. O'Doherty (2010), "States vs. Rewards: Dissociable Neural Prediction Error Signals Underlying Model-based and Model-free Reinforcement Learning," *Neuron* 66, 585-595.

Greenwood, R. and A. Shleifer (2014), "Expectations of Returns and Expected Returns," *Review of Financial Studies* 27, 714-746.

Jin, L. and P. Sui (2022), "Asset Pricing with Return Extrapolation," *Journal of Financial Economics* 145, 273-295.

Khaw, M.W., Li, Z., and M. Woodford (2021), "Cognitive Imprecision and Small-stakes Risk Aversion," *Review of Economic Studies* 88, 1979-2013.

Lee, S., Shimojo, S., and J.P. O'Doherty (2014), "Neural Computations underlying Arbitration between Model-based and Model-free Systems," *Neuron* 81, 687-699.

Liao, J., Peng, C., and N. Zhu (2022), "Extrapolative Bubbles and Trading Volume," *Review of Financial Studies* 35, 1682-1722.

Malmendier, U. and S. Nagel (2011), "Depression Babies: Do Macroeconomic Experiences Affect Risk-taking?" *Quarterly Journal of Economics* 126, 373-416.

Malmendier, U. and J. Wachter (2021), "Memories of Past Experiences and Economic Decisions," Working paper.

McClure, S., Berns, G., and P.R. Montague (2003), "Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum," *Neuron* 38, 339-346.

Montague, P., Dayan, P., and T. Sejnowski (1996), "A Framework for Mesencephalic Dopamine Systems based on Predictive Hebbian Learning," *Journal of Neuroscience* 16, 1936-1947.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and R. Dolan (2003), "Temporal Difference Models and Reward-related Learning in the Human Brain," *Neuron* 38, 329-337.

Pan, W., Su, Z., and J. Yu (2022), "Extrapolative Market Demand," Working paper.

Payzan-LeNestour, E. and P. Bossaerts (2015), "Learning about Unstable, Publicly Unobservable Payoffs," *Review of Financial Studies* 28, 1874-1913.

Schultz, W., Dayan, P., and P.R. Montague (1997), "A Neural Substrate of Prediction and Reward," *Science* 275, 1593-1599.

Shepard, R.N. (1987), "Toward a Universal Law of Generalization for Psychological Science," *Science* 237, 1317-1323.

Sutton R., and A. Barto (2019), *Reinforcement Learning: An Introduction*, MIT Press.

Szepesvari, C. (2010), *Algorithms for Reinforcement Learning.*

Thorndike, E. L. (1933), "A Proof of the Law of Effect," *Science* 77, 173-175.

Tolman, E. C. (1948), "Cognitive Maps in Mice and Men," *Psychological Review* 55, 189-208.

Wachter, J. and M. Kahana (2022), "A Retrieved-context Theory of Financial Decisions," Working paper.

Watkins, C. (1989), "Learning from Delayed Rewards," Ph.D. dissertation, University of Cambridge.

Watkins, C. and P. Dayan (1992), "Q-Learning," *Machine Learning* 8, 279-292.

Woodford, M. (2020), "Modeling Imprecision in Perception, Valuation, and Choice," *Annual Review of Economics* 12, 579-601.