Model-free and Model-based Learning as Joint Drivers of Investor Behavior

Nicholas Barberis and Lawrence Jin

July 2022*

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 link it to a wide range of applications. We show that it provides a foundation for extrapolative demand and experience effects; resolves a puzzling disconnect between investor allocations and beliefs in both the frequency domain and the cross-section; helps explain the dispersion in stock market allocations across investors as well as the inertia in these allocations over time; and sheds light on the persistence of household investment mistakes. 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.

*The authors' affiliations are Yale University and Cornell University, respectively; their e-mails are nick.barberis@yale.edu and lawrence.jin@cornell.edu. We are grateful to Alex Chinco, Cary Frydman, Chen Lian, Ulrike Malmendier, Antonio Rangel, Josh Schwartzstein, Michael Woodford, and seminar participants at Arizona State University, Baruch College, Caltech, Harvard University, Imperial College London, the University of California Los Angeles, the University of Pennsylvania, Yale University, the AFA Annual Meeting, the Behavioral Economics Annual Meeting (BEAM), the Miami Behavioral Finance conference, the NBER Behavioral Finance conference, and the Sloan-Nomis School on Cognitive Foundations of Economic Behavior for very useful feedback. We are also grateful to Colin Camerer, Nathaniel Daw, 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.

1 Introduction

A fundamental question in both economics and psychology asks: How do people choose what actions to take in dynamic settings? The traditional answer in economics is to say that people act "as if" they have solved a dynamic programming problem. Psychologists and neuroscientists, by contrast, have increasingly embraced a different framework for thinking about human 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 model-free and model-based learning algorithms operate in the following dynamic context. At each time, after observing the state of the world, an individual takes an action. In the next period, as a consequence, he receives a reward and arrives in a possibly new state of the world. His goal is to choose an action at each time to maximize the long-term sum of rewards.

The model-free and model-based systems both try to solve this problem by estimating a quantity denoted by $Q(s, a)$, the value of taking action a in state s. However, they do so in different ways. The model-free system is especially different from the framework used by economists in that, as its name indicates, it does not use a model of the world; in other words, it does not use any information about the probabilities of future states and rewards. Instead, it learns from experience. After taking the action a in state s and observing the subsequent reward, it updates its estimate of $Q(s, a)$ by way of two important quantities: a reward prediction error, which, loosely speaking, is the difference between the reward the individual received and the reward he expected; and a learning rate, which controls the extent of the updating. In simple terms, if taking action a in state s leads to a good outcome, the model-free system raises its estimate of $Q(s, a)$ and is therefore more likely to recommend action a if state s is encountered 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]

[1] An early paper on this framework is Daw, Niv, and Dayan (2005). 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, for example, Montague, Dayan, and Sejnowski (1996), Schultz, Dayan, and Montague (1997), McClure, Berns, and Montague (2003), and O'Doherty et al. (2003).

The model-based system is more typical of the frameworks used by economists in that, as in almost all economic models, it makes use of a probability distribution of future rewards and states conditional on past actions and states. There are a number of possible model-based approaches; we use one that is often adopted in research in psychology and that, like the model-free system, has neuroscientific support. In this framework, after taking an action and observing the subsequent reward and state, the model-based system increases the probability it assigns to that reward and state while downweighting the probabilities of other outcomes. To do the updating, it again uses a learning rate and a prediction error, often called a state prediction error, which measures how surprising the realized state and reward are. As with the reward prediction error, there is evidence that the brain computes such state 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 $Q(s, a)$ values produced by each of the model-free and model-based systems and use the resulting "hybrid" $Q(s, a)$ values to make a choice (Glascher et al., 2010; Daw et al., 2011).

In this paper, we import this framework into a simple economic setting, study its properties, and use it to account for a range of facts about investor behavior. The setting is 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.

Our implementation captures the most fundamental difference between the model-free and model-based systems, namely that one uses a model of the environment while the other does not. Another difference between the two systems, also reflected in our framework, is that the model-free system is likely to operate over a more limited time range: because it learns from experienced rewards, it is in operation only when the individual is actively interacting with the environment – for example, only when he is actively experiencing financial markets. By contrast, the model-based system is building a model of how rewards depend on states and actions, and it can do so using data from before the individual started experiencing the environment – for example, from before he started actively investing.

We begin by analyzing the properties of our framework. We focus on 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. We find that, while the model-free algorithm is as simple if not simpler than its model-based counterpart, it has rich implications and leads to new economic intuitions.

We start by looking 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 returns that are positive and that decline for more distant past returns. For many parameter values, the model-free system also recommends an allocation that puts positive and declining weights on past returns. However, relative to the model-based system, it puts substantially more weight on distant past returns. This is because it updates slowly: at each time, it learns primarily about the value of a single action, namely the individual's most recent action. For some parameterizations, it can even put more weight on distant returns than on recent returns. Moreover, the weight it assigns to recent as opposed to distant returns is affected by factors that play no role in the model-based allocation – factors such as the discount rate and the number of allocation choices available to the investor. We also find that the model-free system generates more inertia in investor allocations over time.

We then use our framework to shed light on a range of facts about investor behavior. A prominent idea, motivated by empirical evidence, is that investors have "extrapolative" demand: their demand for a risky asset is a weighted average of the asset's past returns, where the weights are positive and larger for more recent returns. Our analysis 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. Our framework further posits that this demand has two components operating at different frequencies – a model-based component which puts high weight on recent returns and a model-free component which puts substantial weight even on distant past returns.

Our framework also provides a foundation for experience effects – specifically, for the empirical 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 substantial weight on even distant past experienced returns; returns he has not experienced receive less weight. Our framework captures this by way of its model-free component. As noted above, the model-free system puts substantial weight even on distant past experienced returns. However, since it operates only when the individual is experiencing rewards, it puts no weight on returns he has not directly experienced.

Our framework can also resolve a puzzling disconnect between investors' stock market allocations and investors' beliefs. Greenwood and Shleifer (2014), among others, use survey data to show that investor beliefs about future stock market returns depend primarily on recent past market returns. However, Malmendier and Nagel (2011) find that investors' allocations to the stock market depend significantly even on distant past market returns. Two aspects of our framework allow us to reconcile these findings: only the model-based

system has an explicit role for beliefs; and the model-free system proposes allocations that depend even on distant past returns. As a result, when an individual is surveyed about his beliefs regarding future returns, he consults the model-based system and gives an answer that depends primarily on recent past returns. However, when he chooses an allocation, he uses both the model-based and model-free systems and hence chooses an action that depends significantly even on distant past returns. Through a similar mechanism, our framework can also explain the low sensitivity of allocations to beliefs documented by Giglio et al. (2021) in the cross-section of investors.

We show that the framework can also help to account for 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; and the widespread non-participation in the stock market among U.S. households.

Finally, we show that the framework can help explain persistent investment mistakes – in other words, not only why households make suboptimal financial choices, but why they often persist in these choices for many years. In our framework, this behavior stems from the model-free system, and specifically from the fact that this system learns slowly: again, at each moment of time, it learns primarily about the value of a single allocation. As such, it can take a long time to converge to the optimal course of action.

Since the model-free system learns slowly, it is inefficient for an individual to use it to make 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 such, 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 difficult 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 because 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 (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.

In Section 2, we formalize the model-free and model-based learning algorithms and show how they can be applied to a simple portfolio-choice problem. In Section 3, we present an example to show how the two algorithms work and then analyze the properties of our framework. In Section 4, we use the framework to account for a range of facts about investor behavior. Section 5 concludes.

2 Model-free and Model-based Algorithms

Researchers in cognitive psychology and decision 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.

2.1 Psychological background

Under the model-free system, an individual is drawn to actions that have been rewarded in the past. By contrast, under the model-based system, actions are 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 either receives a reward or does not. He repeats this task multiple times with the goal of maximizing the sum of his rewards.

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 independent 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.[3]

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

[3]Charness and Levin (2005) also present an experiment in which model-free and model-based learning – in their terminology, reinforcement learning and Bayesian learning – make different predictions. They, too, find that participant behavior is guided to a significant extent by the model-free system.

to different actions using prediction errors – a reward prediction error in the case of the model-free system and a state prediction error in the case of the model-based system. In an experiment similar to that of Daw et al. (2011), Glascher et al. (2010) use magnetic resonance imaging to show that neural activity in a brain region known as the ventral striatum correlates with the reward prediction error, while neural activity in an area of the prefrontal cortex correlates with the state prediction error. Similar neural evidence has been documented in several other studies.

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 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 the individual takes an action a_t. As a consequence of taking the action a_t in state s_t at time t, 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 denoted $p(s_{t+1}, r_{t+1}|s_t, a_t)$. 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, 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 $p(s_{t+1}, r_{t+1}|a_t, s_t)$ 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 describe how people actually behave, the dynamic programming approach raises an obvious question. It may be difficult to determine the probability distribution $p(\cdot)$; and even if we have a good sense of this distribution, it may be hard, even for professional economists, to then 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. It seems preferable to try to understand individual behavior using a framework that is rooted in, and consistent with, the actual computations the brain performs when making a decision. The framework that we use in this paper has exactly this feature.

We now describe the model-free learning algorithm. 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 to understand decision-making in experimental settings are known as Q-learning and SARSA. In this paper, we use Q-learning. In the Online Appendix, we also consider SARSA and show that it leads to similar predictions.[4]

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 – specifically, 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 $t+1$ 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 learn a good action to take 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 initial estimate of $Q^*(s, a)$ is $Q_t(s, a)$. At

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

time $t+1$, after observing the reward r_{t+1}, its estimate of $Q^*(s,a)$ is updated as

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

where α_t^{MF} is known as the learning rate – the superscript stands for model-free – and 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, $Q_t(s,a)$.

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, often 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 see if this other action has an even higher Q value. In the limit as $\beta \to \infty$, the algorithm chooses the action with the highest 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.

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). Put simply, if an individual 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 individual encounters state s again, he will again choose action a.

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

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

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) 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). However, 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, a large number of studies have found that the brain computes reward prediction errors similar to the one on the right-hand side of equation (4).[5]

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

$$\begin{aligned} Q_{t+1}(s,a) &= Q_t(s,a) + \alpha_{t,+}^{MF}(\text{RPE}) \qquad \text{for RPE} \geq 0 \\ Q_{t+1}(s,a) &= Q_t(s,a) + \alpha_{t,-}^{MF}(\text{RPE}) \qquad \text{for RPE} < 0. \end{aligned} \tag{8}$$

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 (a,s) 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.

[5]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).

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, we first describe the task that we will apply 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.[6] 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. Specifically, 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 simple calculation – see the Appendix – shows that the investor's implied objective is 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

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

a^* of his wealth to the stock market, where

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

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 structure 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 require any information about the structure of the task.

How can Q-learning be applied to the above problem? In principle, we could apply equation (7) 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). While, in the main body of the paper, we put aside state dependence, in the Online Appendix, we re-introduce it and confirm that the key properties of the framework continue to hold.[7]

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

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

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 individual chooses the allocation a and observes the 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

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

where $\alpha_{t,\pm}^{MF}$ equals $\alpha_{t,+}^{MF}$ if the reward prediction error is positive and $\alpha_{t,-}^{MF}$ otherwise. 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 an integral part of how psychologists think about human behavior. By contrast, the term is rarely used in economics or 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 setting, 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.

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.

A large literature in computer science has studied this kind of 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 also engages in generalization (Shepard, 1987). We therefore incorporate generalization into our framework.

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 taking action 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).[8]

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.[9]

[8]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 often 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.

[9]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

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

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, try to build a model of the environment – for example, in our setting, a model 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 the framework we consider, 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 that are sometimes referred to as "state prediction errors." 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 $1 - p_t(R_m = R)$ can be thought of as 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,

topology: the number 70 is closer to 80 than to 20.

$p_t(R_m = R) = 0$, 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 $\alpha_t^{MB} = \alpha^{MB}$ for all t. In this notation, a comma separates a return from its perceived probability, while semicolons separate the different returns:

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

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 state prediction errors analogous to the one in square brackets in (18) (Glascher et al., 2010).[10]

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}, depending on whether the latest net stock market return is positive or negative. Specifically, following the 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

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

[10]While our model-based algorithm is inspired by research in psychology, it is also very 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 the probabilities of all other returns being scaled down by $(1 - \alpha_{t,-}^{MB})$.

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

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

where

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

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 essential difference between a model-free and a model-based system is that the latter makes use of a probability distribution linking future rewards and states to the current action and state – or, in the case without state dependence, a probability distribution linking future rewards to the current action – while the former does not. Our implementation reflects this. Our model-based system has access to a probability distribution for stock market returns and it knows expression (11) linking allocations to portfolio returns. As such, it can construct a distribution of future portfolio returns conditional on some allocation. By contrast, the model-free system has access neither to a probability distribution of stock market returns nor to the relationship between allocations and portfolio returns.

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, while the model-based system, recognizing that choosing B would give more exposure 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 rewarded with a positive net portfolio return. 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. 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 model

An influential framework in psychology posits that people make decisions using a combination of model-free and model-based systems (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" value of $Q(a)$ 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:

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

A well-known hypothesis in psychology is that the value of w varies over time: at each moment, the brain allocates more control to the system that is more certain about the values of different courses of action (Daw, Niv, and Dayan, 2005). We discuss this idea further in Section 4.9. For our main analysis, however, we keep w constant because we find that even this simple case has rich implications.

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 computes Q^{MB} values based on that distribution, as in (22).[11]

3 Properties and Implications

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 $t = 30$.[12,13]

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

[11] 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."

[12] 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 if we change the model frequency.

[13] 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.

possible – the answer would be to use learning rates that decline over time. Specifically, the time t model-based learning rates in (20) and (21) would be[14]

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

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

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

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. For this reason, and because this is the simplest assumption we can make, 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. We use the same baseline parameter values throughout the paper, in part for consistency and in part to show that a single 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 rate γ to 0.97 – this corresponds to an expected investment horizon of 33 years – and the degree of generalization b to 0.0577. At each time, 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 the coarseness of the action set affects the results.

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

As described in Section 2.5, 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 equation (22) (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 an allocation of 70% at date 0.[15]

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.

[15] 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. For the same reason, whenever we consider the model-based system in isolation, we allow for exploration, unless otherwise specified.

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 portfolio 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 an 80% allocation, about the value of a 20% allocation.

3.2 Dependence on past returns

We now analyze a basic property of our framework, one that will be central to several of 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 in particular leads to a rich set of intuitions and implications, some of which are quite distinct from those associated with the model-based system.

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' final 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 $b = 0$.[16] Finally, we set $w = 0.5$, so that the hybrid system puts equal weight on the model-free and

[16]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 will later look at how changing the values of key model parameters affects the results.[17]

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.

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 faster for the model-based system, a property that will play a key role in some of our later 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 for 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 for the model-free system is more subtle, 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: it 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 pulls him toward a high allocation. Why then, on average, does a high return lead to a higher allocation, as 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

[17]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.

when it is 20%. Given that this mechanism is less direct than the one for the model-based system, it is natural that, as shown in Figure 2, the dependence of the allocation on recent stock market returns is quantitatively smaller for the model-free system.

The weights that the model-based system puts on past returns decline as we go further into the past. Mathematically, this is because every time the model-based system updates its perceived return distribution, it scales down the probabilities of past returns by a proportional factor, reducing the importance of these earlier returns. 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 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 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? Whenever 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; in short, it learns slowly. 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 regardless of the learning rate, the model-free allocation puts positive and declining weights on past returns, with the decline being slightly more pronounced for higher learning rates.

For many parameter values, including those used in Figures 2 and 3, the model-free allocation puts more weight on recent than distant past returns. However, the model-free system can exhibit richer behavior than this. For example, it sometimes puts more weight on distant than on recent past returns. Moreover, for the model-free system, the relationship between allocations and past returns is affected by factors that play no role for the model-based system.

To illustrate this, each of the four graphs in Figure 4 varies a key model parameter while keeping the other parameters at their benchmark levels. The top-left graph in Figure 4 plots the coefficients in a regression of the model-free allocation on past stock market returns for four values of the generalization parameter b: 0, 0.0577, 0.115, and 0.23. The first of these

values corresponds to no generalization; the other three values give the Gaussian function in (17), normalized as a probability distribution, a standard deviation equal to that of a uniform distribution with width 0.2, 0.4, and 0.8, respectively.

The figure shows something striking: as we raise the degree of generalization, we begin to see an increasing relationship between allocations and past returns, so that the model-free allocation puts more relative weight on *distant* past returns. To see the intuition, suppose that, when he first enters financial markets, an investor chooses an allocation of 80% and that the stock market then performs well. For a high degree of generalization, as with $b = 0.23$, this immediately creates a cluster of allocations ranging from, say, 60% to 100%, with high Q values. This makes it likely that the investor will keep choosing an allocation in this range for a long time to come, thereby giving the early returns he encountered an outsize influence on his later allocations.

The top-right graph in Figure 4 plots the relationship between the model-free allocation and past returns for three different values of β, which controls the degree of exploration, namely 10, 50, and 500. Recall that, as β rises, the investor explores less: he is more likely to choose the allocation with the highest estimated Q value. We find that, for a wide range of values of β – any β below 100 – the model-free allocation puts positive and declining weights on past returns, as it does for our benchmark case of $\beta = 30$. However, when β is very high – higher than 100 – we begin to see an increasing relationship between allocations and past returns, at least over some range. To see why, suppose that, soon after the investor enters financial markets, the stock market posts a high return, raising the Q value of his most recent allocation. If the value of β is high, the investor is likely to stick with this allocation for a substantial period of time. As such, the early returns he experiences have a large effect on his subsequent allocations.

The bottom-left graph plots the relationship between the model-free allocation and past returns for three different values of the discount rate γ, namely 0.3, 0.9, and 0.99. As we lower γ, the allocation puts much greater weight on recent past returns. This is a striking result in that it links an investor's expected future investment horizon to the relative weight he puts on recent as opposed to distant past returns when choosing an allocation. For the model-based system, by contrast, the discount rate does not affect the dependence of allocations on past returns.

Thus far, we have allowed investors to select from one of 11 possible allocations. The bottom-right graph in Figure 4 shows how the time T allocation depends on past returns as we vary the number of allocation options available to investors, ranging from three, namely {0%, 50%, 100%}, up to 21, namely {0%, 5%, ... , 95%, 100%}. The graph shows that, as we

lower the number of possible allocations, the relationship between the time T allocation and past returns, while initially downward-sloping, becomes much flatter, thereby giving distant past returns a larger relative role. This property of the model-free system again distinguishes it sharply from the model-based system, where the number of possible allocations has little impact on the relationship between the time T allocation and past returns.

One way of understanding the bottom-right graph is to note that reducing the number of allocation options is akin to increasing the degree of generalization: since generalization leads the investor to treat nearby allocations in a similar way, a large number of allocations coupled with generalization is like a small number of allocations without generalization. Just as in the top-left graph we see a flat or increasing relationship between the time T allocation and returns for high levels of generalization, so in the lower-right graph, we see a flat and, in places, increasing relationship for a lower number of allocation choices.[18]

In summary, the model-free system has rich implications for the relationship between allocations and past returns. While this relationship is typically downward-sloping, it is sometimes upward-sloping. Moreover, there is structure to this relationship: we know the conditions under which it is more likely to be downward- rather than upward-sloping. Finally, the relationship between model-free allocations and past returns is affected by factors that play little to no role in the model-based system. We return to some of these novel implications in Section 4.

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, for example, about the dependence of model-free allocations on past returns. Nonetheless, for certain special cases, we *are* able to derive such results. We present these results and their proofs in Online Appendix A. This analysis confirms a fundamental property we have emphasized in this section, namely that, relative to the model-based system, the model-free system tends to put significantly more weight on distant past returns.

3.3 Dispersion and variability in allocations

We now consider some other properties of the model-free and model-based systems – properties related to the dispersion and variability in investor allocations. By "dispersion," we

[18]The results in Figures 2 to 4 are for an annual-frequency implementation of our framework. We have studied the effect of changing the model frequency. If we fix the learning rates α^{MB} and α^{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.

mean the standard deviation, across investors, of their date T allocations. By "variability," we mean the standard deviation of investors' allocations over time: for each investor in turn, we compute the standard deviation of his allocations over time – the standard deviation of $\{a_{T-j}\}_{j=1}^{30}$ for this investor – and then average these standard deviations across investors. We obtain two results. First, the variability in investor allocations is typically lower under the model-free system: under this system, there is more "inertia" in an investor's allocations from period to period. Second, under the model-free system, there is more dispersion in investors' final allocations.

To demonstrate these results, we now allow for dispersion in learning rates across investors.[19] Specifically, 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 $\sigma = 0.2$. As in Section 3.1, we have $b = 0.0577$, so that there is some generalization, and we set the new parameter Δ to 0.5. We take 1,000 investors, expose all of them to the same sequence of stock market returns from $t = -L$ to $t = T$, and compute the resulting dispersion and variability. We repeat this exercise 300 times for different return sequences and average the resulting set of dispersion and variability estimates.

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 three model 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 main finding is that the dashed lines are substantially below the solid lines: the model-free system leads to lower variability than the model-based system. To understand this, note that, under the model-based system, investors tend to increase their allocation following a good return and lower their allocation following a poor return; as a consequence, there is substantial variability. By contrast, under the model-free system, an investor can become "stuck" at a particular allocation: if, early on, the investor chooses some allocation to the stock market and the market then performs well, the Q value of the chosen allocation will be pushed up, raising the chance that the investor will keep choosing this allocation in subsequent years. The three upper graphs show that the difference in variability levels between the two systems is increasing in the mean learning rate and decreasing in the

[19] 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.

amount of exploration and the dispersion in learning rates.

The solid and dashed lines in the three lower graphs in Figure 5 plot the dispersion in final allocations across investors for the model-based and model-free systems, respectively, as we vary β, $\bar{\alpha}$, and Δ, while keeping the other parameter values at their benchmark levels. While the difference between the two systems is not as stark as in the case of variability, the graphs show that dispersion in allocations is typically higher for the model-free system. To understand this, note that, under the model-based system, following a high stock market return, all investors perceive an improvement in the distribution of stock market returns and hence tend to raise their allocation to the stock market; this, in turn, keeps the dispersion in allocations across investors at a relatively low level. The model-free system, by contrast, generates higher dispersion. This stems from the interaction of the probabilistic action choice and the reinforcement inherent in this system. At time 0, the probabilistic action choice in (14) leads to dispersed allocations across investors. If the stock market then performs well, this reinforces each investor's initial allocation, leading each investor to persist with his initial allocation and preserving the dispersion in allocations across investors.

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 shed light on real-world financial behavior. Second, one component of the framework is, by definition, "model-free": it 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 start by showing that a simple parameterization of the framework can qualitatively, and even quantitatively, address a range of facts about investor behavior. 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 model 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, making 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 will formally estimate the value of w that best fits the data.

We now use our framework to address a range of applications.

4.1 Extrapolative demand

The first application follows directly from the analysis of Section 3.2, but it is an important one that merits further discussion. 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.[20]

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 that depends positively on past returns and more so on recent past returns. To be clear, the mechanism in the case of the model-based system is not particularly novel; however, for the model-free system, the mechanism is new to the finance literature. We explained it in full in Section 3.2. In brief: 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; he is therefore more likely to 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 its past returns, with weights on past returns 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 in that, relative to the analysis of 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 present 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 the existing finance literature 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; indeed, we saw in Section 3.2 that, in some cases, the model-free system puts more weight on distant than on recent returns. As such, while empirical analyses suggest that the allocations of real-world investors put more weight on recent returns, this may mask a model-free component that puts more weight on distant returns but is outweighed by a model-based component

[20] 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 et al. (2015, 2018), Cassella and Gulen (2018), Jin and Sui (2021), Liao, Peng, and Zhu (2021), and Pan, Su, and Yu (2021).

that puts heavy weight on recent returns. We make use of this two-component structure of extrapolative demand in subsequent applications.

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 offers 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 hope to replicate. The more important one is that, if an investor begins participating in financial markets at time t, his allocation 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 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. As a way of capturing both features, Malmendier and Nagel (2011) propose that investors' decisions are based on a weighted average of past returns in which, for an investor with n years of experience, the weight on the return k years ago is

$$(n-k)^{\lambda}/A, \tag{28}$$

where λ is estimated to be approximately 1.5 and A is a normalization factor, and where the weight on returns the investor did not experience is zero.

To see whether 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 (28). The dashed line plots the functional form in (28) for the cohort in question, and the vertical dotted line marks the point at which the cohort enters financial markets.

By comparing, within each graph, the solid and dashed lines, 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 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 an experienced return that is 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 to a greater extent.

Similar patterns can be seen in the other five 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.

Using an analogous approach to that described above, our framework can also capture several other types of experience effects in financial markets – for example, that after experiencing good returns on their investments in a particular industry, IPO stock, or lottery-type stock, people are more likely to purchase another stock in that industry, another IPO stock, or another lottery-type stock, respectively (Kaustia and Knupfer, 2008; Huang, 2019; Hui et al., 2021).

4.3 Investor beliefs and the frequency disconnect

Several studies have found that investor beliefs about future stock market returns are a positive function of recent past stock market returns. Our framework can capture this; but more strikingly, it can also help explain 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 discuss in the next section.

The disconnect in the frequency domain is simple to state. While studies of investor expectations about future returns find that these expectations depend heavily on *recent* past

returns, investor stock market allocations depend to a substantial extent even on distant past returns (Malmendier and Nagel, 2011).[21]

Two features of our framework allow it to explain this disconnect. First, while each investor's allocation is based on both the model-free and model-based systems, only one of these – 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 an investor's beliefs, which are generated by the model-based system, will put heavy weight on recent returns, while his allocations, which are based on both systems, will put a greater relative weight on distant past returns. As such, the framework 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 have been 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 have been 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 correlated with flows into equity market mutual funds. We stress that our framework is consistent with such findings: in our simulated data, there is a strong 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 shed light on.

[21] We can formalize this in the following way. When Malmendier and Nagel (2011) use the weights in (28) to characterize the relationship between an investor's allocation and the past returns he has experienced, they obtain an estimate of $\lambda \approx 1.5$. Suppose that we now take the functional form in (28) 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 $\lambda \approx 50$, which puts much greater weight on recent returns.

4.4 Investor beliefs and the cross-sectional disconnect

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

Our framework can help capture this disconnect. The mechanism is similar to that for the frequency disconnect and 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 less weight on recent returns.

We now examine this quantitatively. Table 2 reports, for three different values of the weight w on the model-based system, the coefficient in a regression of investors' stock market allocations at time T on their expected returns on the stock market over the next year. The table shows that our framework can help explain the cross-sectional disconnect described above: for our benchmark value of $w = 0.5$, the regression coefficient in our simulated data, 1.25, is similar to that obtained by Giglio et al. (2021) in actual data. Moreover, the table shows that the model-free system plays an important role in this result: as we increase the weight on the model-free system, the sensitivity of allocations to beliefs falls.

4.5 Dispersion and inertia in household allocations

Households differ in their asset allocations: some participate in the stock market, while others do not; and among households that participate, the fraction of wealth they invest in the stock market varies substantially. It is not easy to explain these differing allocations: regressions of allocations on explanatory variables have a low R-squared.

The framework of this paper offers two ways of thinking about this dispersion in holdings. First, it says that these differences are due in part to differences in investor learning rates. In

our simulated data, investors with higher values of $\alpha_+^{MF} - \alpha_-^{MF}$ and $\alpha_+^{MB} - \alpha_-^{MB}$ have higher stock market allocations, on average: these investors update more in response to positive returns or reward prediction errors, which, in turn, tends to raise the Q values of higher stock market allocations.

A second, more novel, possibility is one discussed earlier in connection with Figure 5. The lower-right panel in that figure shows that the model-free system generates substantial dispersion in investor allocations at time T even when all investors have the *same* learning rates. The dispersion here is driven by the interaction of the probabilistic action choice and model-free reinforcement. If, as a result of the probabilistic choice, investor A chooses a low allocation to the stock market early on while investor B chooses a high allocation, and the stock market then posts a high return, choosing a low (high) allocation will be reinforced for investor A (B), leading to persistent differences in these investors' allocations.

While there is substantial cross-sectional dispersion in households' allocations to the stock market, there is also individual-level inertia in these allocations over time. 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. 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. If, after an investor chooses some allocation to the stock market, the market posts a good return, the Q value of that allocation goes up substantially, which makes it more likely that the investor will keep choosing that allocation in the future.

4.6 Non-participation

For the 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; the traditional Expected Utility model, by contrast, predicts that all investors will allocate at least some fraction of their wealth to the stock market. 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 raises the probability that, in a subsequent period, he will switch to a 0% allocation to the

market. Importantly, once he does so, the model-free system will update only the Q value of the 0% allocation: since, generalization aside, it learns only about the action taken, it stops learning about the stock market, and, in particular, fails to learn that the stock market has better properties than indicated by the 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 that better suits this particular application. In this version, there are 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? It is not important that the stock market allocation is a 100% allocation; we obtain similar results if the two allocations are 0% and 50%, say.[22] 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 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 0 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. Under the model-free system, 43% of investors spend at least 80% of the 30 years not participating in the stock market, in other words, at a 0% allocation. By contrast, under the model-based system, fewer than 1% of investors spend more than 80% of the 30 years 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 17% of investors have a streak of this length.

More interestingly, the simulated data support the mechanism for non-participation we 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.

[22] One possibility is that the investor uses a separate model-free / model-based framework for each of two decisions: a two-allocation framework for the participation decision, and a framework with more possible allocations to decide on his allocation conditional on participation.

4.7 Persistent investment mistakes

Many households make suboptimal financial choices; moreover, they often persist in these choices for long periods of time. The framework of this paper 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 the 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. The gross return on asset i, R_i, is distributed as

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

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 are allocating their wealth to the best option, asset 10?

We answer this question separately for a rational, model-based benchmark and for a model-free system. In the case of the model-based benchmark, all investors use declining learning rates analogous 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. We also set $\beta = \infty$, so that there is no exploration; exploration adds no value when the model-based system operates in isolation. In the case of the model-free system, 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 $\Delta = 0.5$. 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 is 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 B, we present the full updating equations for the two systems.

We find that, in the case of the model-based system, at time $T = 30$, 47% of investors are allocating to the best option, asset 10. By contrast, for the model-free system, at time $T = 30$, just 21% of investors are allocating to asset 10. Consistent with our claim above, then, the model-free system learns slowly: it takes longer to figure out the sensible course

of action. This result does not hinge on the constant learning rate. If investors instead use the model-free system in conjunction with a declining learning rate – one that satisfies the conditions for long-run convergence of Q values in (27) – then, at time $T = 30$, just 19% of investors are allocating to asset 10.[23]

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.

4.8 Parameter estimation

Throughout this section, we have taken a simple parameterization of our framework and shown that it can provide a qualitative and approximate quantitative match to a number of facts about investor behavior. We now do an estimation exercise to see which parameter values best match the data. Our empirical targets are two concrete and central facts from earlier in this section, namely the experience effects of Malmendier and Nagel (2011) and the sensitivity of allocations to beliefs from Giglio et al. (2021). 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 do the estimation in two steps. First, we use data on investor beliefs to estimate $\bar{\alpha}^{MB}$. With this in hand, we then estimate $\bar{\alpha}^{MF}$, β, and w by targeting the empirical facts on experience effects and the allocation-belief sensitivity. 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 $b = 0.0577$.

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 returns. We take Gallup data from October 1996 to November 2011 on average beliefs about future stock market returns and regress these beliefs on past annual stock market returns. The coefficient on the past year's return is 0.127, and the coefficient on the return two years in the past is 0.037; the ratio of the second coefficient to the first is 0.29. We search for a value of $\bar{\alpha}_{MB}$ that, in simulated data, best matches the first coefficient, 0.127, and the rate of decline in the coefficients, 0.29; intuitively, we are trying to match the level and slope of the relationship between beliefs and returns. To do this, we take $30,000$ investors in six

[23]Specifically, we use the learning rate $1/(1 + t^{0.6})$, which satisfies the conditions in (27).

cohorts of $5,000$ each; each investor sees a different sequence of stock market returns from time $t = -L$ to time $t = T$. For a given value of $\bar{\alpha}^{MB}$, we draw each investor's model-based learning rates from a uniform distribution centered at $\bar{\alpha}_{MB}$ and with width $\Delta = 0.5$. We then compute investors' beliefs at each time, as determined by the model-based system and by equations (20) and (21) in particular. Finally, we regress investors' beliefs at time T on the past 30 years of stock market returns they have been exposed to, and record the coefficient c_1 on the most recent annual return and the coefficient c_2 on the second most recent annual return. We repeat this exercise for many different values of $\bar{\alpha}^{MB}$ and select the value of $\bar{\alpha}^{MB}$ that minimizes

$$(c_1 - 0.127)^2 + (\frac{c_2}{c_1} - 0.29)^2. \tag{29}$$

We find this to be $\bar{\alpha}^{MB} = 0.38$.

With this value of $\bar{\alpha}^{MB}$ in hand, 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 1 in the data. For given values of $\bar{\alpha}^{MF}$, β, and w, we can compute this coefficient, d, in our simulated data.

Our second target is the functional form in (28), with $\lambda = 1.5$, which Malmendier and Nagel (2011) use to capture empirical experience effects. Intuitively, we are looking for parameter values that minimize the distance between the red and blue lines in the six graphs in Figure 7. Specifically, for given values of $\bar{\alpha}^{MF}$, β, and w, and for cohort 1, we run a regression in our simulated data of the time T allocations on the past 30 years of returns; we then compute the L^2 norm of the difference between the vector of the 30 coefficients (these correspond to the blue line in the top-left graph in Figure 7) and the vector of 30 values implied by (28) (the red line in the graph). We call this MSE_1, the mean-squared error for cohort 1. In a similar way, we compute MSE_i for $i = 2$ to 6, which correspond to cohorts 2 through 6.

We repeat the above exercise for many different values of $\{\bar{\alpha}^{MF}, \beta, w\}$. In other words, for many values of $\{\bar{\alpha}^{MF}, \beta, w\}$, we compute the quantity

$$\sum_{i=1}^{6} \text{MSE}_i + (d-1)^2 \tag{30}$$

and identify the parameter values that minimize this quantity. The first term in (30) targets the empirical data on experience effects, while the second term targets the empirical sensitivity of allocations to beliefs.

We find that the parameter values that minimize (30) are $\bar{\alpha}^{MF} = 0.66$, $\beta = 20$, and $w = 0.46$. The value of w is particularly well identified. The reason is the following. In the first term in (30), we are trying to match the empirical pattern of experience effects. As shown by the red lines in Figure 7, this involves both an initial sharp 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 initial sharp 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 sharp decline. As such, to match both features of the data, we need to put substantial weight on both systems – as it turns out, a roughly equal weight on the two systems.

4.9 Extensions

We now discuss some possible extensions of our framework.

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 or dramatic news. Such an assumption can be incorporated into our framework and may lead to useful new predictions.

Time-varying weights on the two systems. We have taken w, the weight on the model-based system, to be constant over time. A well-known hypothesis in psychology is that w varies over time: the brain assigns more control to the system that is currently more certain about the values of different courses of action (Daw, Niv, and Dayan, 2005). For example, when a person first interacts with a new environment, w may take a high value: the model-based system learns quickly and is therefore more useful. Over time, as the model-free system accumulates more experience, the brain may assign it more control, lowering w. In our framework, this would predict that the stock market allocations of older people will react less to recent returns and will exhibit more inertia over time.

Other model-based frameworks. When we specify the model-free system in Section 2, we do not have much flexibility: 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. Indeed, in Online Appendix C, we replace with Q-learning with SARSA, an alternative model-free framework, and show that it leads to similar predictions. By contrast, when specifying the

model-based part of our framework, we have a wider range of choices. In Section 2, we adopted a model-based system inspired by those used in psychology, but others are 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 $Q(s, a)$ and show that even this simple case has many applications. In Online Appendix D, 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. We leave richer analyses of state dependence 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 investors also use the model-*free* system to make inferences about beliefs. When an investor is asked for his beliefs about the stock market's future return or risk, it is natural that he will first 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 where return and risk are positively related.

5 Conclusion

When economists try to explain behavior in dynamic settings, they usually assume that people act "as if" they have solved a dynamic programming problem. By contrast, psychologists and neuroscientists are increasingly embracing a different framework, 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 link it to a range of applications. We show that it provides a foundation for extrapolative demand and experience effects; resolves a puzzling disconnect between investor allocations and beliefs in both the frequency domain and the cross-section; can help explain the dispersion across investors in their stock market allocations as well as the inertia in these allocations over time; and can shed light on why many households make persistent investment mistakes. Overall, our results suggest that model-free reinforcement learning, which has had only a small footprint in economics until now, may be more useful to economists than previously thought.

There are two broad directions for future research. We can apply the framework proposed here to other economic domains – for example, to think about consumption choice. We can also incorporate richer psychological assumptions – for example, about time-varying learning rates or weights on the model-free system, or about state dependence. We expect that both of these broad directions will be fruitful in shedding light on economic data.

6 Appendix

A portfolio-choice problem that fits the model-free / model-based learning framework (Section 2.3)

The investor's objective is to maximize

$$(1-\gamma)E(\log W_1) + \gamma(1-\gamma)E(\log W_2) + \gamma^2(1-\gamma)E(\log W_3) + \dots \tag{31}$$

where

$$W_t = W_0 \Pi_{\tau=1}^{t} R_{p,\tau} \tag{32}$$

is his wealth at time t. Substituting (32) into (31) and rearranging, the objective function becomes

$$\log W_0 + E\sum_{t=1}^{\infty} \gamma^{t-1} \log R_{p,t},$$

as in (10).

7 References

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.

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

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.

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.

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

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

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.

Huang, X. (2018), "Mark Twain's Cat: Investment Experience, Categorical Thinking, and Stock Selection," *Journal of Financial Economics* 131, 404-432.

Hui, C., Liu, Y-J., Xu, X., and J. Yu (2021), "Priming and Stock Preferences: Evidence from IPO Lotteries," Working paper.

Jin, L. and P. Sui (2021), "Asset Pricing with Return Extrapolation," *Journal of Financial Economics*, forthcoming.

Kaustia, M. and S. Knupfer (2008), "Do Investors Overweight Personal Experience? Evidence from IPO Subscriptions," *Journal of Finance* 63, 2679-2702.

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 (2021), "Extrapolative Bubbles and Trading Volume," *Review of Financial Studies*, forthcoming.

Malmendier, U. and S. Nagel (2011), "Depression Babies: Do Macroeconomic Experiences

Affect Risk-taking?" *Quarterly Journal of Economics* 126, 373-416.

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 (2021), "Extrapolative Market Demand," Working paper, Tsinghua University.

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.

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.

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 $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $b = 0.0577$, $\mu = 0.01$, and $\sigma = 0.2$.

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

Table 2. Sensitivity of investor allocations to investor beliefs. The table reports the sensitivity of investors' stock market allocations a_T at time T to their time T expectations of the future one-year stock market return for various values of the weight w on the model-based system. There are 300,000 investors: six cohorts of 50,000 investors each which enter financial markets at different dates. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

w	Sensitivity
0.2	0.7
0.5	1.25
1	1.91

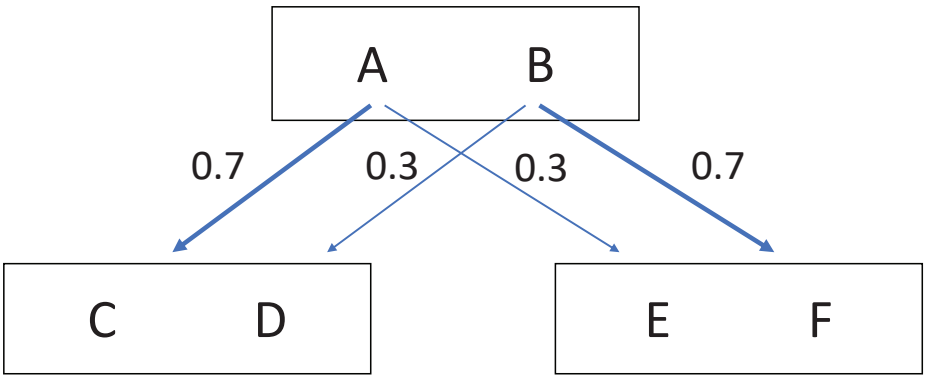

Figure 1. The diagram shows the structure of an experiment in Daw et al. (2011). In the first stage, the participant has a choice between two options, A and B; in the second stage, he chooses either between options C and D or between options E and F. The arrows indicate the transition probabilities from the first to the second stage. After making a choice at the second stage, the participant either receives a reward or does not.

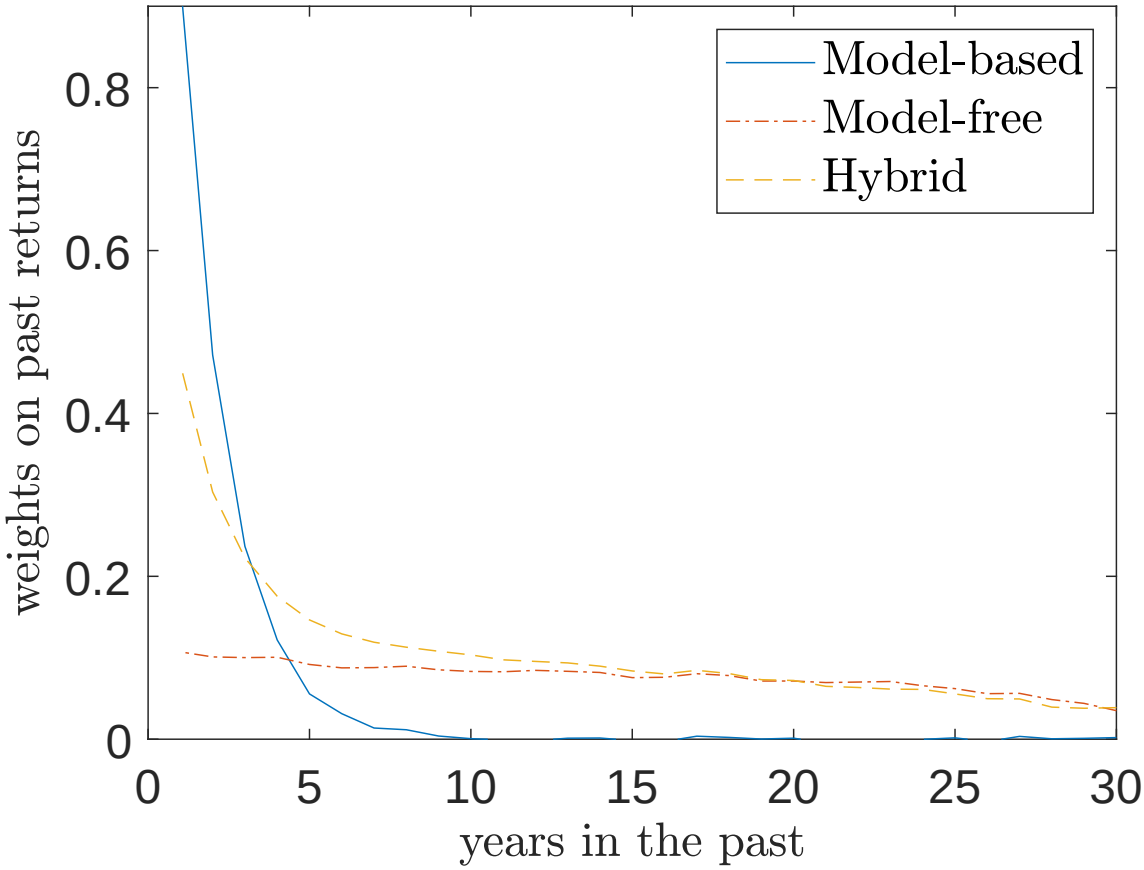

Figure 2. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors were exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system. The 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.

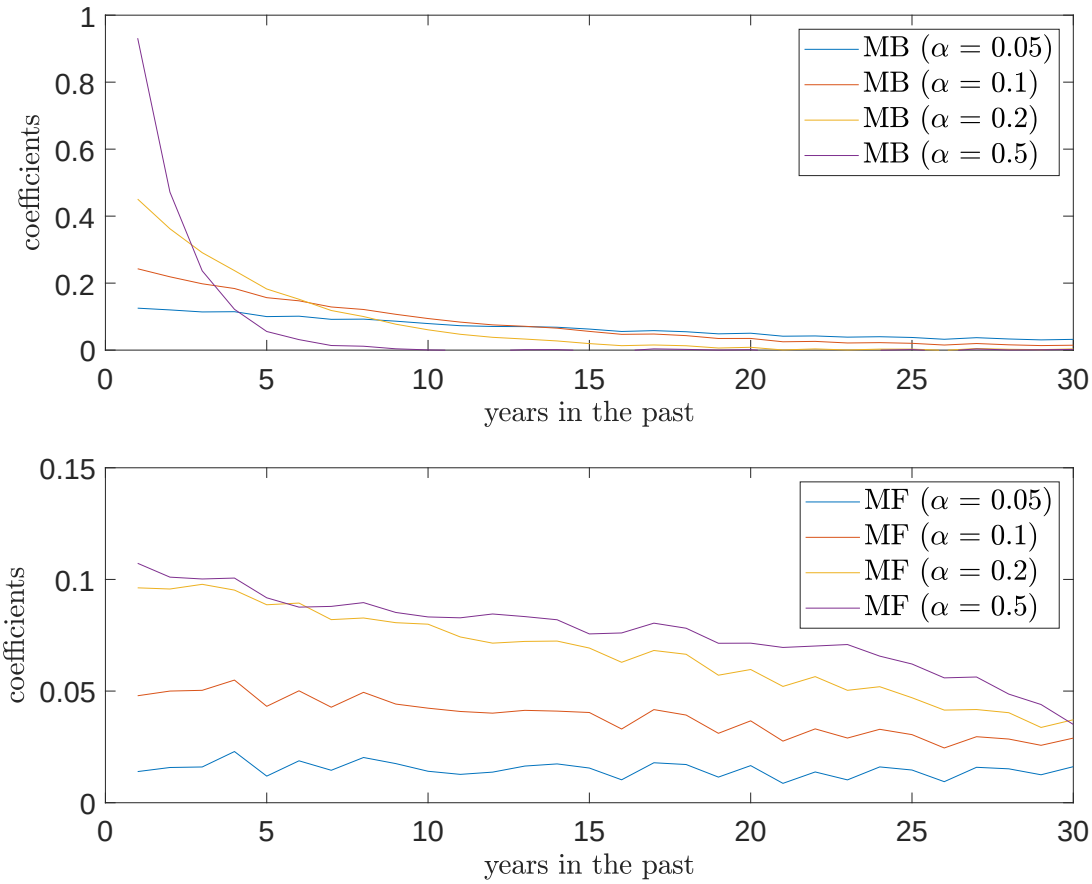

Figure 3. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors were exposed to. The top graph plots the coefficients for the model-based system for four values of the learning rates α_+^{MB} and α_-^{MB}, namely 0.05 (blue), 0.1 (red), 0.2 (yellow), and 0.5 (magenta). 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 (blue), 0.1 (red), 0.2 (yellow), and 0.5 (magenta). 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.

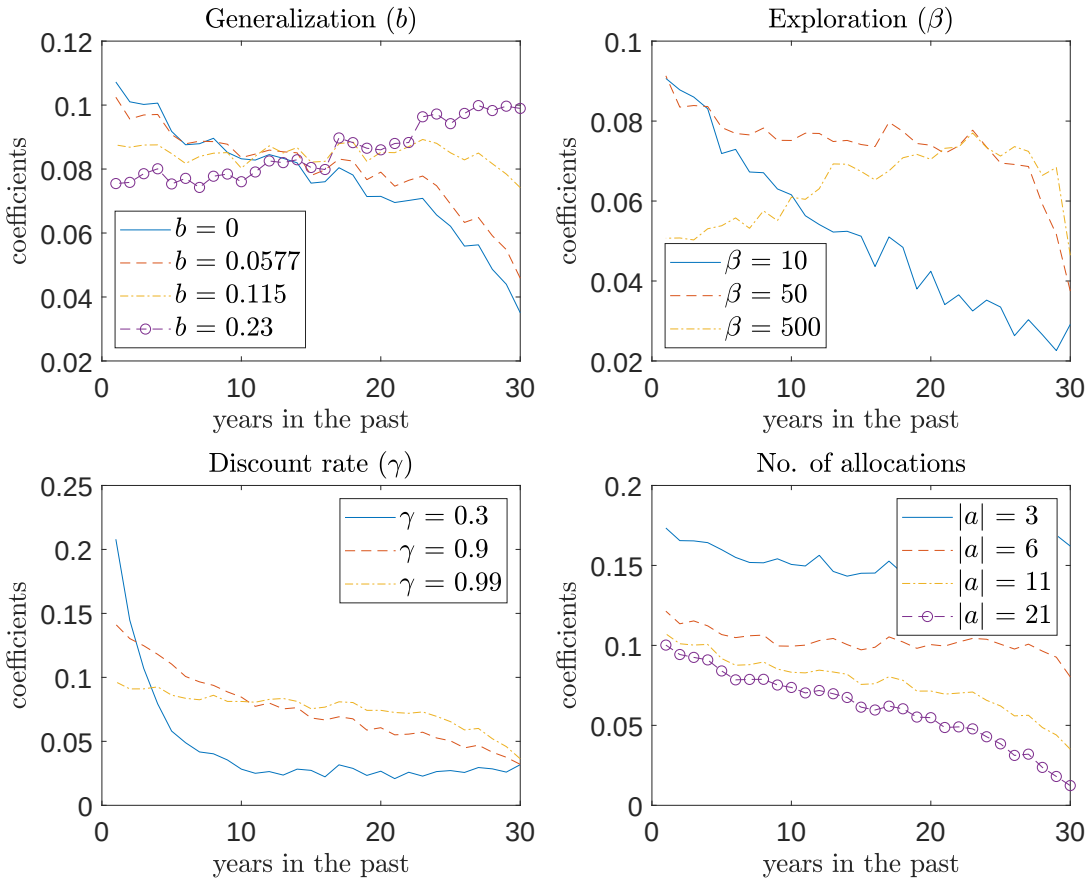

Figure 4. For different sets of parameter values, we run a regression of investors' allocations to the stock market a_T at time T under the model-free system on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors were exposed to and plot the coefficients. The lines in the top-left, top-right, bottom-left, and bottom-right graphs correspond, respectively, to four values of the generalization parameter b, namely 0 (blue), 0.0577 (red), 0.115 (yellow), and 0.23 (magenta); to three values of the exploration parameter β, namely 10 (blue), 50 (red), and 500 (yellow); to three values of the discount rate γ, namely 0.3 (blue), 0.9 (red), and 0.99 (yellow); and to different numbers of allocation choices, namely 3 (blue), 6 (red), 11 (yellow), and 21 (magenta). There are 300,000 investors. For the remaining parameters, we set $L = T = 30$, $\alpha_{\pm}^{MF} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

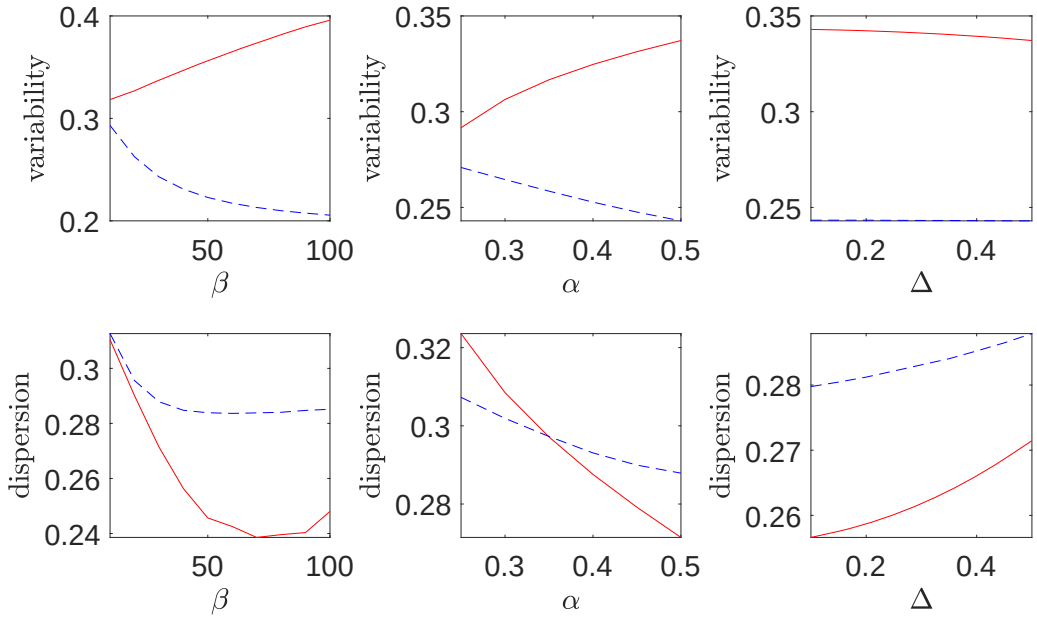

Figure 5. The upper graphs plot the variability of stock market allocations – the standard deviation of allocations between time 0 and time T, computed for each investor in turn and averaged across investors. The lower graphs plot the dispersion, across investors, of their stock market allocations at 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 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 $b = 0.0577$.

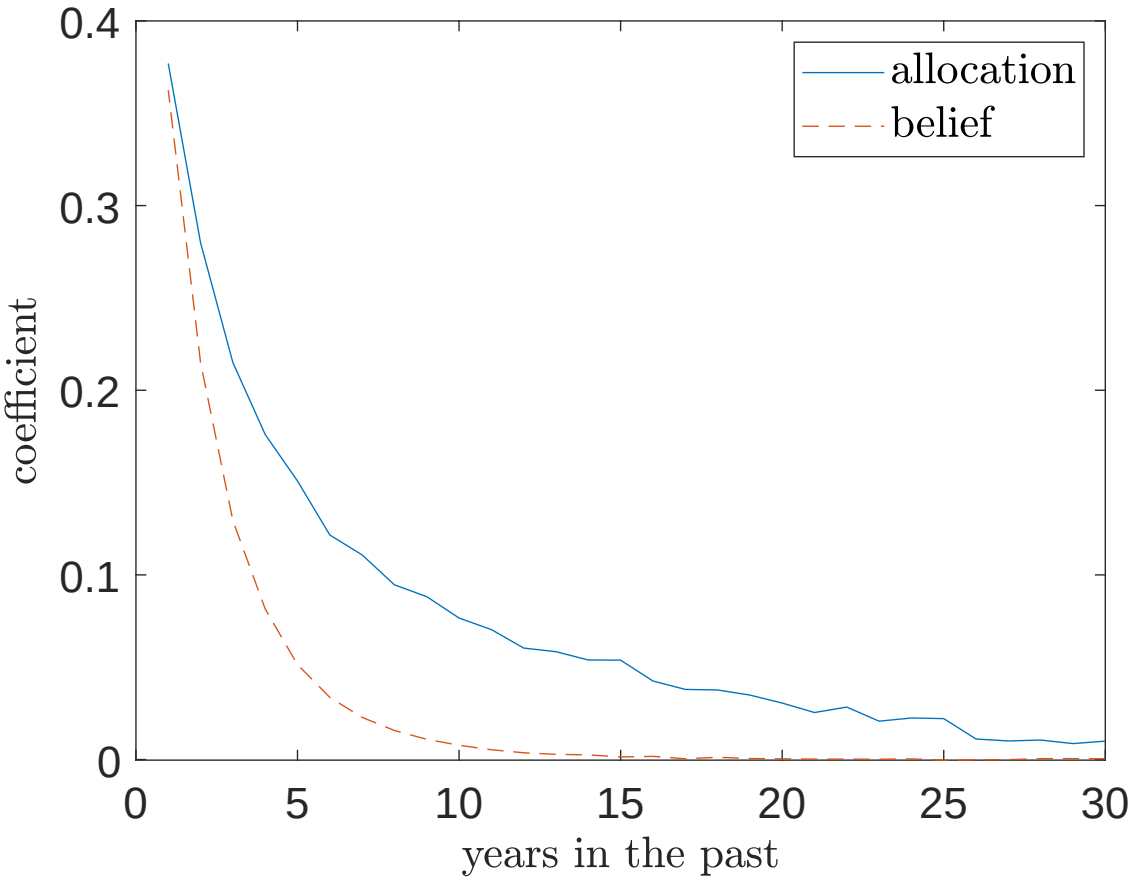

Figure 6. The solid line plots the coefficients in a regression of the stock market allocation a_T at date T chosen by investors who use a hybrid system to make decisions on the past 30 years of stock market returns the investors were exposed to. The dashed line plots the coefficients in a regression 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 $w = 0.5$.

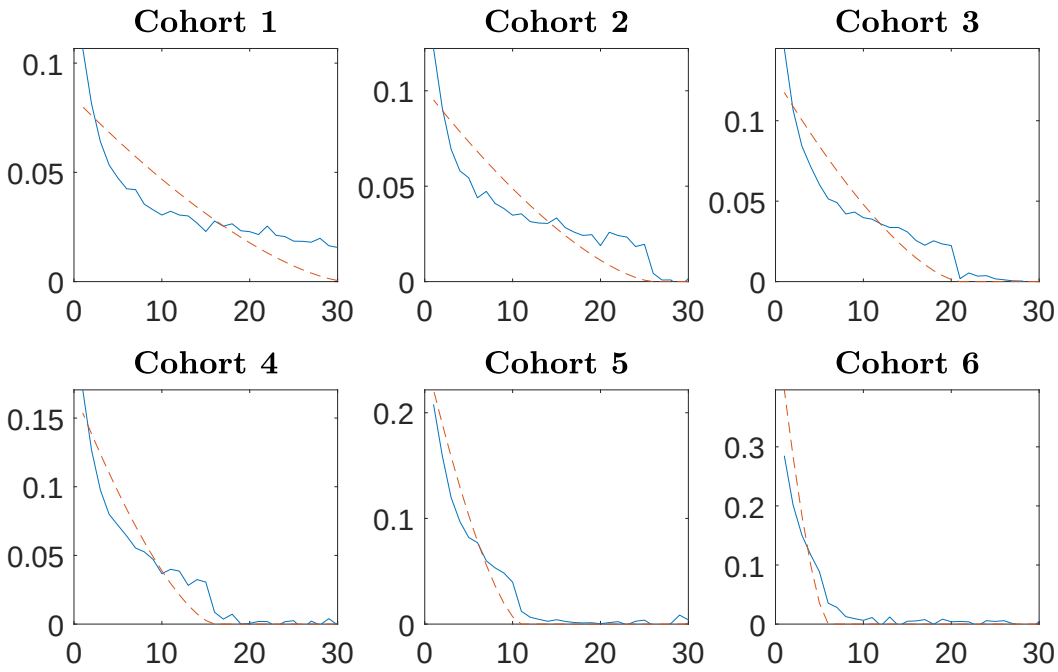

Figure 7. The six graphs correspond to six cohorts of investors. In each graph, the solid line plots the coefficients, normalized to sum to one, in a regression of the time T stock market allocations a_T of the investors in that cohort on the past 30 years of stock market returns they were exposed to. 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 $(n-k)^\lambda/A$, where k is the number of years in the past, $\lambda = 1.5$, and A is a normalizing constant.

ONLINE APPENDIX

A. Analytical Results

While the Q-learning algorithm is simple to state, it is difficult to derive analytical results about its predictions. Nonetheless, in some cases, we *are* able to derive such results – specifically about how the stock market allocation it recommends depends on past market returns. In this section, we present these results and their proofs.

We start with the case in which the learning rates $\alpha = 1$, the discount rate $\gamma = 0$, and there are just two possible allocations, namely $a = 0$ and $a = 1$. For model-free and model-based learning, respectively, Theorems 1 and 2 below present analytical results on how the allocation recommended by each system depends on past returns. By comparing equations (1) and (10), we confirm that the model-free system puts substantially greater weight on distant past returns.

We then turn to a less restrictive case where the learning rates α can take any value between 0 and 1; once again, $\gamma = 0$ and there are two possible allocations. For the model-free and model-based algorithms, respectively, Theorems 3 and 4 below present analytical results about the dependence of the recommended allocation on past returns. Comparing equations (13) and (26)-(27), we again see that the model-free algorithm puts substantially greater weight on distant past returns.

We have also been able to prove analytical results in the case where the discount rate γ is greater than zero. However, the resulting expressions are much messier and do not provide much additional intuition.

Theorem 1 (Model-free learning): Assume that $\alpha = 1$, $\beta > 0$, $\gamma = 0$, $R_f = 1$, and that there are two possible allocations $\{0, 1\}$. Set $Q_0(0) = Q_0(1) = 0$. Further assume that $R_{m,t} \equiv R$ for all periods $t \geq 1$.

Given these assumptions, the following result is true:

$$\lim_{t\to\infty} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \frac{\beta R^{2\beta-1}}{(R^\beta + 1)^{k+3}} \tag{1}$$

for $k \geq 0$.

Proof: At any time $t > 0$,

$$\begin{aligned} Q_t(0) &= \log(R_f) = 1, \\ Q_t(1) &= \log(R_{m,t'}), \end{aligned} \tag{2}$$

where t' is the most recent time such that $a_{t'-1} = 1$ and $R_{m,t'}$ is the market return from time $t' - 1$ to time t'.

Equation (2) allows us to express the expected allocation $\mathbb{E}[a_t]$ as

$$\begin{aligned}
\mathbb{E}[a_t] &= \mathbb{P}(a_t = 1) \\
&= \sum_{i=0}^{t-1} \mathbb{P}(a_t = 1 | i \text{ is the largest index s.t. } a_i = 1) \times \mathbb{P}(a_i = 1) \\
&\quad + \mathbb{P}(a_t = 1 | a_0 = \ldots = a_{t-1} = 0) \times \mathbb{P}(a_0 = \ldots = a_{t-1} = 0) \\
&= \left(\sum_{i=0}^{t-1} \frac{R_{m,i+1}^{\beta}}{R_{m,i+1}^{\beta} + 1} \left(\frac{1}{R_{m,i+1}^{\beta} + 1} \right)^{t-i-1} \times \mathbb{P}(a_i = 1) \right) + \frac{1}{2^{t+1}}. \qquad (3)
\end{aligned}$$

Given the assumption that $R_{m,t} \equiv R$ for all periods $t \geq 1$, we conjecture and then verify the following result:

$$\mathbb{P}(a_t = 1) = \frac{(2^{t+1} - 1)R^{\beta} + 1}{2^{t+1}(R^{\beta} + 1)}, \quad \forall t \geq 0. \qquad (4)$$

The verification of (4) is as follows. When $t = 0$, equation (4) implies that $\mathbb{P}(a_0 = 1) = \frac{1}{2}$, which is clearly true. For $t = j \geq 1$, suppose (4) is true for $0 \leq i \leq j - 1$. Then, we have, from equation (3),

$$\begin{aligned}
\mathbb{P}(a_j = 1) &= \left(\sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{j-i}} \times \mathbb{P}(a_i = 1) \right) + \frac{1}{2^{j+1}} \\
&= \left(\sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{j-i}} \times \frac{(2^{i+1} - 1)R^{\beta} + 1}{2^{i+1}(R^{\beta} + 1)} \right) + \frac{1}{2^{j+1}} \\
&= \frac{R^{\beta}(1 - 2^{-j})}{R^{\beta} + 1} + \frac{1}{2^{j+1}} = \frac{(2^{j+1} - 1)R^{\beta} + 1}{2^{j+1}(R^{\beta} + 1)}. \qquad (5)
\end{aligned}$$

That is, (4) is also true for $t = j$.

Equation (4) allows us to derive $\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}}$, the sensitivity of the expected allocation to past returns. We first consider the case with $k = 0$. In this case,

$$\begin{aligned}
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t}} &= \frac{\partial \mathbb{P}(a_t = 1)}{\partial R_{m,t}} = \frac{\partial \left[\frac{R_{m,t}^{\beta}}{R_{m,t}^{\beta} + 1} \mathbb{P}(a_{t-1} = 1) \right]}{\partial R_{m,t}} \\
&= \frac{\beta R_{m,t}^{\beta - 1}}{(R_{m,t}^{\beta} + 1)^2} \mathbb{P}(a_{t-1} = 1) = \frac{\beta R^{\beta - 1}}{(R^{\beta} + 1)^2} \frac{(2^t - 1)R^{\beta} + 1}{2^t(R^{\beta} + 1)}. \qquad (6)
\end{aligned}$$

As t goes to infinity, we obtain

$$\lim_{t \to \infty} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t}} = \frac{\beta R^{2\beta - 1}}{(R^{\beta} + 1)^3}, \qquad (7)$$

which is the same as (1) when $k = 0$.

Next, we consider the case with $k > 0$. In this case,

$$\begin{aligned}
\frac{\partial \mathbb{P}(a_t = 1)}{\partial R_{m,t-k}} &= \left(\sum_{i=t-k}^{t-1} \frac{R_{m,i+1}^{\beta}}{(R_{m,i+1}^{\beta} + 1)^{t-i}} \cdot \frac{\partial \mathbb{P}(a_i = 1)}{\partial R_{m,t-k}} \right) + \frac{\partial \left[\frac{R_{m,t-k}^{\beta}}{(R_{m,t-k}^{\beta}+1)^{k+1}} \mathbb{P}(a_{t-k-1} = 1) \right]}{\partial R_{m,t-k}} \\
&= \left(\sum_{i=t-k}^{t-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{t-i}} \cdot \frac{\partial \mathbb{P}(a_i = 1)}{\partial R_{m,t-k}} \right) + \frac{\beta R^{\beta-1} - k\beta R^{2\beta-1}}{(R^{\beta} + 1)^{k+2}} \cdot \mathbb{P}(a_{t-k-1} = 1) \\
&= \sum_{i=0}^{k-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{i+1}} \cdot \frac{\partial \mathbb{P}(a_{t-i-1} = 1)}{\partial R_{m,t-k}} \\
&\quad + \frac{\beta R^{\beta-1} - k\beta R^{2\beta-1}}{(R^{\beta} + 1)^{k+2}} \cdot \frac{(2^{t-k} - 1)R^{\beta} + 1}{2^{t-k}(R^{\beta} + 1)}. \qquad (8)
\end{aligned}$$

Suppose (1) is true for $0 \leq k \leq j - 1$, then

$$\begin{aligned}
\lim_{t\to\infty} \frac{\partial \mathbb{P}(a_t = 1)}{\partial R_{m,t-j}} &= \sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{i+1}} \cdot \lim_{t\to\infty} \frac{\partial \mathbb{P}(a_{t-i-1} = 1)}{\partial R_{m,t-j}} \\
&\quad + \frac{\beta R^{\beta-1} - j\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j+2}} \cdot \frac{R^{\beta}}{R^{\beta} + 1} \\
&= \left(\sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{i+1}} \cdot \frac{\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j-i+2}} \right) + \frac{\beta R^{\beta-1} - j\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j+2}} \cdot \frac{R^{\beta}}{R^{\beta} + 1} \\
&= \frac{j\beta R^{3\beta-1}}{(R^{\beta} + 1)^{j+3}} + \frac{\beta R^{2\beta-1} - j\beta R^{3\beta-1}}{(R^{\beta} + 1)^{j+3}} = \frac{\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j+3}}. \qquad (9)
\end{aligned}$$

That is, (1) holds for $k = j$. Equation (9) completes an inductive proof of (1). ■

Theorem 2 (Model-based learning): Assume that $\alpha = 1$, $\beta > 0$, $\gamma = 0$, $R_f = 1$, and that there are two possible allocations $\{0, 1\}$. Set $Q_0(0) = Q_0(1) = 0$.

Given these assumptions, the following result is true:

$$\begin{aligned}
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t}} &= \frac{\beta R_{m,t}^{\beta-1}}{(R_{m,t}^{\beta} + 1)^2}, \\
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} &= 0, \quad k > 0. \qquad (10)
\end{aligned}$$

Proof: At any time $t > 0$,

$$\begin{aligned}
Q_t(0) &= 0, \\
Q_t(1) &= \log(R_{m,t}). \qquad (11)
\end{aligned}$$

The softmax rule implies

$$\mathbb{E}[a_t] = \mathbb{P}(a_t = 1) = \frac{R_{m,t}^{\beta}}{R_{m,t}^{\beta} + 1}. \tag{12}$$

Taking the derivative of (12) with respect to $R_{m,t-k}$ leads to (10). ■

Theorem 3 (Model-free learning): Assume that $\alpha \in (0, 1]$, $\gamma = 0$, $\beta > 0$, $R_f = 1$, and that there are two possible allocations $\{0, 1\}$. Set $Q_0(0) = Q_0(1) = 0$. Assume that $R_{m,i} \equiv R$ for all periods $i \geq 1$. Further assume that, when investors invest in the stock market for the first time, the learning rate in the Q-learning algorithm is 1; all the subsequent learning rates are set to α.

Given these assumptions, the following result is true:

$$\lim_{t\to\infty} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{2\beta-1}}{(R^{\beta}+1)^3} \left(\frac{R^{\beta} + 1 - \alpha R^{\beta}}{R^{\beta} + 1} \right)^k. \tag{13}$$

Proof: Let $[t]$ denote $\{0, 1, \ldots, t\}$ and $[j, t]$ denote $\{j, j + 1, \ldots, t\}$. Then, by definition,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t} \frac{\partial \left[\mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1]) \mathbb{P}(a_i = b_i, \forall i \in [t-1]) \right]}{\partial R_{t-k}} \tag{14}$$

$$= \sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t} \frac{\partial \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} \mathbb{P}(a_i = b_i, \forall i \in [t-1]) \tag{15}$$

$$+ \sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t} \frac{\partial \mathbb{P}(a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1]). \tag{16}$$

We analyze the expressions in (15) and (16) separately. First, we derive $\lim_{t\to\infty}(15)$, the limit of the expression in (15) as t goes to infinity. We have

$$\frac{\partial \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} = \frac{\partial \left(\frac{e^{\beta Q_t(1)}}{e^{\beta Q_t(1)}+1} \right)}{\partial R_{t-k}} = \frac{1}{(e^{\beta Q_t(1)} + 1)^2} \frac{\partial e^{\beta Q_t(1)}}{\partial R_{t-k}}. \tag{17}$$

If $b_{t-k-1} = 0$, then R_{t-k} is never used to update the Q values; as such, $\frac{\partial \mathbb{P}(a_t=1|a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}} = 0$. If, on the other hand, $b_{t-k-1} = 1$, then we note $\frac{1}{(e^{\beta Q_t(1)}+1)^2} = \frac{1}{(R^{\beta}+1)^2}$, because the Q value for a 100% allocation to the stock market gets updated to $\log(R)$ when investors invest in the stock market for the first time and then stays at $\log(R)$ afterwards.

To further derive $\frac{\partial e^{\beta Q_t(1)}}{\partial R_{t-k}}$ in (17), we let n denote the number of indices i, with $i \in \{t - k, \ldots, t - 1\}$ and $b_i = 1$. We then proceed by considering two cases. The first case is when $b_0 = b_1 = \ldots = b_{t-k-2} = 0$. In this case, $Q_t(1)$ can be written as the sum of $(1 - \alpha)^n \log(R_{t-k})$ and a term unrelated to R_{t-k}. As such,

$$\frac{\partial e^{\beta Q_t(1)}}{\partial R_{t-k}} = \frac{(1-\alpha)^n \beta e^{\beta Q_t(1)}}{R} = (1-\alpha)^n \beta R^{\beta-1} \tag{18}$$

and (17) can be simplified as

$$\frac{\partial \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} = \frac{(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta + 1)^2}. \tag{19}$$

The second case is when $b_0, \ldots, b_{t-k-2}$ are not all equal to zero. In this case, $Q_t(1)$ can be written as the sum of $\alpha(1-\alpha)^n \log(R_{t-k})$ and a term unrelated to R_{t-k}. As such, (17) can be simplified as

$$\frac{\partial \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} = \frac{\alpha(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta + 1)^2}. \tag{20}$$

Substituting (19) and (20) back into (15), we obtain

$$\begin{aligned}
(15) \;=\; & \sum_{n=0}^{k} \sum_{\substack{(b_{t-k},\ldots,b_{t-1}) \in (0,1)^k \\ \sum_{j=t-k}^{j=t-1} b_j = n,\, b_{t-k-1}=1}} \frac{(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta+1)^2} \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) = (0,\ldots,0)}\right) \\
& + \sum_{n=0}^{k} \sum_{\substack{(b_{t-k},\ldots,b_{t-1}) \in (0,1)^k \\ \sum_{j=t-k}^{j=t-1} b_j = n,\, b_{t-k-1}=1}} \frac{\alpha(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta+1)^2} \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) \neq (0,\ldots,0)}\right).
\end{aligned} \tag{21}$$

Note that

$$0 \leq \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) = (0,\ldots,0)}\right) \leq \mathbb{P}((a_0, \ldots, a_{t-k-2}) = (0, \ldots, 0)) = \frac{1}{2^{t-k-1}}. \tag{22}$$

Therefore $\lim_{t\to\infty} \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) = (0,\ldots,0)}\right) = 0$ and $\lim_{t\to\infty} \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) \neq (0,\ldots,0)}\right) = \lim_{t\to\infty} \mathbb{P}(a_i = b_i, \forall i \in [t-k-1, t-1])$. Also note that

$$\begin{aligned}
\mathbb{P}(a_t = 1) \;=\; & \mathbb{P}(a_t = 1 | (a_0, \ldots, a_{t-1}) = (0, \ldots, 0)) \cdot \mathbb{P}((a_0, \ldots, a_{t-1}) = (0, \ldots, 0)) \\
& + \mathbb{P}(a_t = 1 | (a_0, \ldots, a_{t-1}) \neq (0, \ldots, 0)) \cdot \mathbb{P}((a_0, \ldots, a_{t-1}) \neq (0, \ldots, 0)) \\
\;=\; & \frac{1}{2}\left(\frac{1}{2}\right)^t + \frac{R^\beta}{R^\beta + 1}\left(1 - \left(\frac{1}{2}\right)^t\right),
\end{aligned} \tag{23}$$

which means $\lim_{t\to\infty}\mathbb{P}(a_t=1)=\frac{R^\beta}{R^\beta+1}$. These limiting results further imply

$$
\begin{aligned}
&\lim_{t\to\infty}(15)\\
=\;&\sum_{n=0}^{k}\frac{\alpha(1-\alpha)^n\beta R^{\beta-1}}{(R^\beta+1)^2}\lim_{t\to\infty}\sum_{\substack{(b_{t-k},\ldots,b_{t-1})\in(0,1)^k\\ \sum_{j=t-k}^{j=t-1}b_j=n,\,b_{t-k-1}=1}}\mathbb{P}\big(\substack{a_i=b_i,\forall i\in[t-k-1,t-1],\\(a_0,\ldots,a_{t-k-2})\neq(0,\ldots,0)}\big)\\
=\;&\sum_{n=0}^{k}\frac{\alpha(1-\alpha)^n\beta R^{\beta-1}}{(R^\beta+1)^2}\Big(\lim_{t\to\infty}\mathbb{P}(a_{t-k-1}=1)\Big)\lim_{t\to\infty}\sum_{\substack{(b_{t-k},\ldots,b_{t-1})\in(0,1)^k\\ \sum_{j=t-k}^{j=t-1}b_j=n}}\mathbb{P}\big(a_i=b_i,\forall i\in[t-k,t-1]|a_{t-k-1}=1\big)\\
=\;&\sum_{n=0}^{k}\frac{\alpha(1-\alpha)^n\beta R^{\beta-1}}{(R^\beta+1)^2}\frac{R^\beta}{R^\beta+1}\binom{k}{n}\left(\frac{R^\beta}{R^\beta+1}\right)^n\left(\frac{1}{R^\beta+1}\right)^{k-n}\\
=\;&\frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^{3+k}}\sum_{n=0}^{k}\binom{k}{n}(1-\alpha)^nR^{n\beta}\\
=\;&\frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^{3+k}}(1+(1-\alpha)R^\beta)^k=\frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^3}\left(\frac{R^\beta+1-\alpha R^\beta}{R^\beta+1}\right)^k. \qquad (24)
\end{aligned}
$$

We now turn to (16). We have

$$
\begin{aligned}
(16)\;=\;&\sum_{\substack{(b_0,\ldots,b_{t-1})\in\{0,1\}^t\\(b_0,\ldots,b_{t-1})\neq(0,\ldots,0)}}\frac{\partial\mathbb{P}(a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}}\frac{R^\beta}{R^\beta+1}\\
&+\frac{\mathbb{P}((a_0,\ldots,a_{t-1})=(0,\ldots,0))}{\partial R_{t-k}}\cdot\frac{1}{2}\\
=\;&\sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t}\frac{\partial\mathbb{P}(a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}}\frac{R^\beta}{R^\beta+1}\\
&+\frac{\mathbb{P}((a_0,\ldots,a_{t-1})=(0,\ldots,0))}{\partial R_{t-k}}\left(\frac{1}{2}-\frac{R^\beta}{R^\beta+1}\right)\\
=\;&\frac{\partial\sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t}\mathbb{P}(a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}}\frac{R^\beta}{R^\beta+1}\\
=\;&0. \qquad (25)
\end{aligned}
$$

Finally, (24) and (25) together lead to (13). ■

Theorem 4 (Model-based learning): Assume that $\alpha\in(0,1]$, $\gamma=0$, $\beta>0$, $R_f=1$, and that there are two possible allocations $\{0,\,1\}$. Set $Q_0(0)=Q_0(1)=0$. Assume that $R_{m,i}\equiv R$ for all periods $i\geq 1$.

Given these assumptions,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{\beta-1}}{(R^\beta+1)^2}(1-\alpha)^k \tag{26}$$

for $0 \leq k < t-1$. For $k = t-1$,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,1}} = \frac{\beta R^{\beta-1}}{(R^\beta+1)^2}(1-\alpha)^{t-1}. \tag{27}$$

Proof: For $t \geq 1$, we have

$$\begin{aligned} Q_t(1) - Q_t(0) &= \mathbb{E}_t^p(R_{m,t+1}) \\ &= (1-\alpha)^{t-1}\log(R_{m,1}) + \alpha\sum_{j=2}^{t}(1-\alpha)^{t-j}R_{m,j} \\ &= \log(R). \end{aligned} \tag{28}$$

For $0 \leq k < t-1$,

$$\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,t-k}} = \frac{\alpha(1-\alpha)^k}{R}, \tag{29}$$

and for $k = t-1$,

$$\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,1}} = \frac{(1-\alpha)^{t-1}}{R}. \tag{30}$$

We can express $\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}}$ as follows

$$\begin{aligned} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} &= \frac{\partial\left(\frac{e^{\beta(Q_t(1)-Q_t(0))}}{e^{\beta(Q_t(1)-Q_t(0))}+1}\right)}{\partial R_{m,t-k}} \\ &= \frac{\beta e^{\beta(Q_t(1)-Q_t(0))}}{(e^{\beta(Q_t(1)-Q_t(0))}+1)^2}\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,t-k}} \\ &= \frac{\beta R^\beta}{(R^\beta+1)^2}\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,t-k}}. \end{aligned} \tag{31}$$

Substituting (29) and (30) into (31) then gives (26) and (27), respectively. ■

B. Updating Equations for a Multi-Asset Setting

In this section, we provide the model-free and model-based updating equations for the setting described in Section 4.7, one with ten risky assets.

For the setting of Section 4.7, model-free and model-based learning operate in an analogous way to what is described in Section 2 for the case of one risky asset. An investor's action a_t at time t is drawn from the set $\{1,\ldots,n\}$, where action $a_t = i$ means that the investor allocates his wealth to asset i at time t. If, at time t, the investor takes action i,

then, at time $t+1$, the model-free value of this action is updated as

$$Q_{t+1}^{MF}(i) = Q_t^{MF}(i) + \alpha_{t,\pm}^{MF}[\log R_{p,t+1} + \gamma \max_{j\epsilon\{1,\ldots,n\}} Q_t^{MF}(j) - Q_t^{MF}(i)],$$

which is analogous to equation (13) in the main text. Meanwhile, after observing the return R from asset i at time $t+1$, the model-based system uses the update equation

$$p_{t+1}(R_i = R) = \alpha_t^{MB} = \frac{1}{t+1},$$

analogous to (20), (21), and (26) in the main text, to create a perceived return distribution for each asset. Its model-based estimate of the Q value of each action is given by

$$\begin{aligned} Q_t(i) &= E_t^p(\log R_i) + \frac{\gamma}{1-\gamma} E_t^p(\log R_j) \\ j &= \arg\max_l E_t^p(\log R_l), \end{aligned}$$

where the expectation E_t^p is taken under the investor's time t perceived distribution of each asset's returns.

C. SARSA: An Alternative Model-free Framework

The model-free frameworks most widely used by psychologists are Q-learning and SARSA. In the main part of the paper, we focus on Q-learning. In this section, we consider SARSA instead. In particular, we examine how the stock market allocation recommended by SARSA depends on past market returns. We find that the results for SARSA are similar to those for Q-learning: relative to model-based learning, SARSA and Q-learning both put substantially more weight on distant past market returns.

We first describe how SARSA works. Suppose that the action space is $\mathbf{a} = \{a_1, a_2, \ldots, a_N\}$. At time 0, all Q values are set to zero: $Q_0^{MF}(a) = 0$, $\forall a$. The investor chooses one of the N allocations with equal probability; we denote this initial allocation by a_0. At each subsequent time t, the investor observes the portfolio return $R_{p,t}$ generated by the stock market return $R_{m,t}$ and by a_{t-1}, his time $t-1$ allocation. He then chooses his allocation a_t probabilistically, according to

$$\mathbb{P}(a_t = a) = \frac{\exp[\beta Q_t^{MF}(a)]}{\sum_{a'} \exp[\beta Q_t^{MF}(a')]}, \tag{32}$$

and given $R_{p,t}$ and a_t, he replaces $Q_t^{MF}(a_{t-1})$, the Q value for his previous allocation, by

$$Q_t^{MF}(a_{t-1}) + \alpha_{t,\pm}^{MF}\left[\log R_{p,t} + \gamma Q_t^{MF}(a_t) - Q_t^{MF}(a_{t-1})\right]. \tag{33}$$

Analogous to the analysis in Section 3.2, we examine how investors' date T allocation a_T recommended by each of SARSA, Q-learning, and model-based learning depends on the past market returns investors have been exposed to. Figure A1 presents the results and leads to two observations. First, for SARSA and Q-learning, the weights the allocation a_T puts on past stock market returns are quantitatively similar. The only exception is the weight on the

most recent stock market return: in the case of SARSA, the allocation a_T is determined by Q values that do not depend on the most recent return $R_{m,T}$; this allocation therefore puts zero weight on $R_{m,T}$. Second, the allocation recommended by model-based learning depends primarily on recent past returns; by contrast, the allocations recommended by Q-learning and SARSA depend significantly even on distant past returns.

D. Allowing for State Dependence

In the main text, we focus on the case with no state dependence – we find that this case already delivers rich results. In this section, we incorporate an explicit state dependence into our framework, both to show how this can be done and to check that it does not affect a fundamental property of the framework we rely on in Section 4, namely that the model-free system puts substantially more weight on distant past returns.

We consider an exogenous state s_t with the transition matrix

$$\begin{array}{c} \\ s_t = h \\ s_t = l \end{array} \begin{array}{c} \begin{array}{cc} s_{t+1} = h & s_{t+1} = l \end{array} \\ \begin{pmatrix} 1-\chi & \chi \\ \chi & 1-\chi \end{pmatrix} \end{array}, \tag{34}$$

where $0 < \chi < 1$ represents the probability of a transition between h and l. Suppose that in state h, $\log R_{m,t} = \mu_h + \sigma\varepsilon_t$; and that in state l, $\log R_{m,t} = \mu_l + \sigma\varepsilon_t$. Here $\varepsilon_t \sim N(0,1)$ is $i.i.d$ over time, and $\mu_h > \mu_l$.

For model-free learning, the Q-learning algorithm is

$$Q^{MF}_{t+1}(a, s_t) = Q^{MF}_t(a, s_t) + \alpha^{MF}_{t,\pm}\left[\log R_{p,t+1} + \gamma \max_{a'} Q^{MF}_t(a', s_t) - Q^{MF}_t(a, s_t)\right] \tag{35}$$

with $s_t \in \{h, l\}$.

For model-based learning, the probability estimates are updated according to

$$\mathbb{P}^{\text{new}}(R_m = R, s_t) = \mathbb{P}^{\text{old}}(R_m = R, s_t) + \alpha^{MB}_{t,\pm}[1 - \mathbb{P}^{\text{new}}(R_m = R, s_t)] \tag{36}$$

with $s_t \in \{h, l\}$. The model-based Q values are

$$\begin{aligned} Q^{MB}_t(a, s_t = h) &= \mathbb{E}^{p,h}_t[\log((1-a)R_f + aR_{m,t+1})] + \gamma[(1-\chi)V^h + \chi V^l], \\ Q^{MB}_t(a, s_t = l) &= \mathbb{E}^{p,l}_t[\log((1-a)R_f + aR_{m,t+1})] + \gamma[(1-\chi)V^l + \chi V^h], \end{aligned} \tag{37}$$

where the superscript "p, h" denotes the time t perceived return distribution conditional on state h, and the superscript "p, l" denotes the time t perceived return distribution conditional on state l. The two values V^h and V^l represent the optimal valuations given state h or state l. They are defined by

$$\begin{aligned} V^h &= \mathbb{E}^{p,h}_t[\log((1-a^*_h)R_f + a^*_h R_{m,t+1})] + \gamma[(1-\chi)V^h + \chi V^l], \\ V^l &= \mathbb{E}^{p,l}_t[\log((1-a^*_l)R_f + a^*_l R_{m,t+1})] + \gamma[(1-\chi)V^l + \chi V^h], \end{aligned} \tag{38}$$

where $a^*_h = \arg\max_a \mathbb{E}^{p,h}_t[\log((1-a)R_f + aR_{m,t+1})]$ and $a^*_l = \arg\max_a \mathbb{E}^{p,l}_t[\log((1-a)R_f +$

$aR_{m,t+1})]$. Solving (38) for V^h and V^l gives

$$\begin{aligned} V^h &= \frac{(1-\gamma+\gamma\chi)\mathbb{E}_t^{p,h}[\log((1-a_h^*)R_f + a_h^* R_{m,t+1})] + \gamma\chi\mathbb{E}_t^{p,l}[\log((1-a_l^*)R_f + a_l^* R_{m,t+1})]}{(1-\gamma)(1-\gamma+2\gamma\chi)}, \\ V^l &= \frac{\gamma\chi\mathbb{E}_t^{p,h}[\log((1-a_h^*)R_f + a_h^* R_{m,t+1})] + (1-\gamma+\gamma\chi)\mathbb{E}_t^{p,l}[\log((1-a_l^*)R_f + a_l^* R_{m,t+1})]}{(1-\gamma)(1-\gamma+2\gamma\chi)}. \end{aligned} \tag{39}$$

The hybrid Q values are

$$Q_t^{HYB}(a, s_t) = (1-w)Q_t^{MF}(a, s_t) + wQ_t^{MB}(a, s_t). \tag{40}$$

Finally, actions are chosen probabilistically according to

$$\mathbb{P}(a_t = a, s_t) = \frac{\exp[\beta Q_t^{HYB}(a, s_t)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(a', s_t)]}. \tag{41}$$

We now examine the implications of this state-dependent model by way of a numerical example. We set $\mu_h = 0.03$, $\mu_l = -0.01$, and $\chi = 0.25$. The values of all the other parameters are the same as in the main text. We consider two cases. In the first case, investors recognize that there are two states; they update the model-free and model-based Q values according to (35) and (37). In the second case, investors fail to recognize the two states; they assume that there is only one state and use the single-state models described in the main text.

For both cases, we examine how investors' date T allocations a_T depend on the past market returns investors have been exposed to. Figure A2 presents the results for the case where investors recognize the two states; Figure A3 presents the results for the case where investors assume there to be only one state. Comparing Figures A2 and A3 shows that incorporating a state structure into our framework does not alter our finding in the main text about the way allocations depend on past returns: for both model-free and model-based learning, the allocations put weights on past stock market returns that are positive and that decline the further back we go into the past; importantly, the decline is much faster in the case of model-based learning.

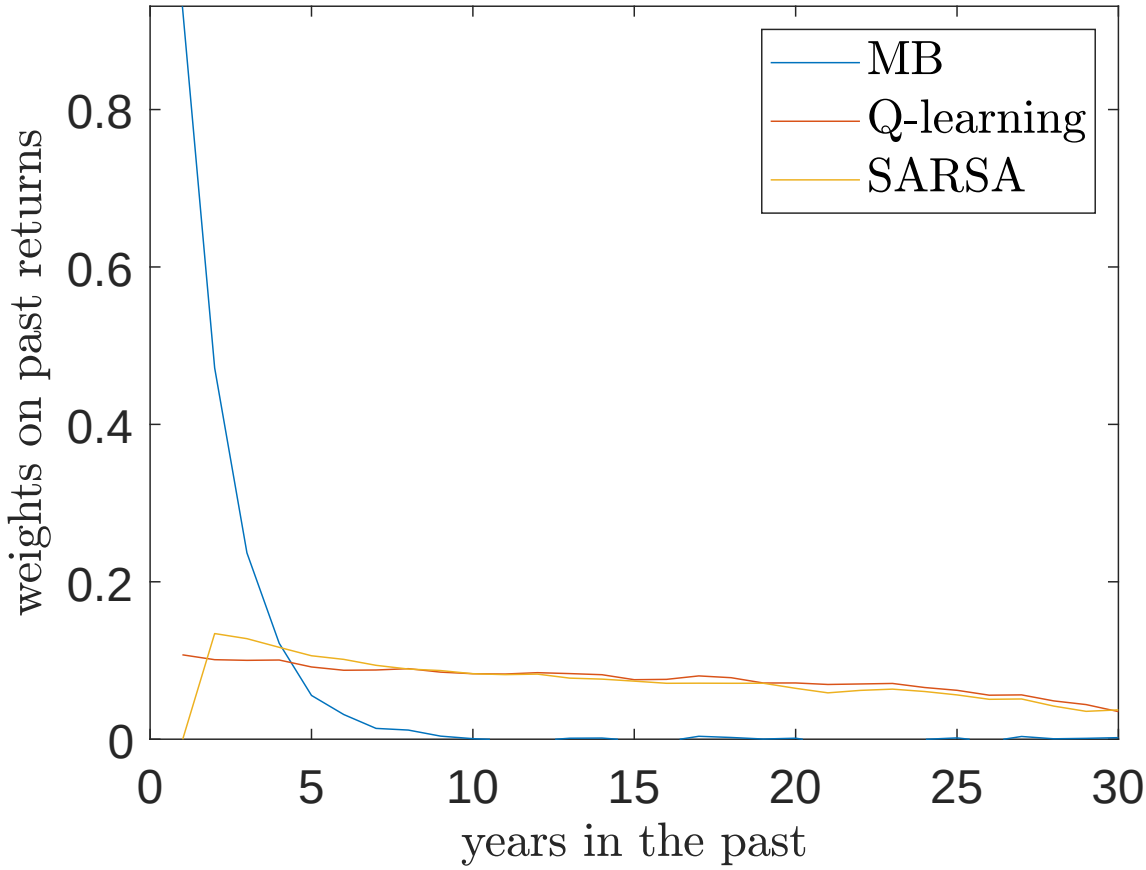

Figure A1. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors were exposed to and plot the coefficients for three cases: model-based learning; model-free Q-learning; and model-free SARSA. There are 300,000 investors. We set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

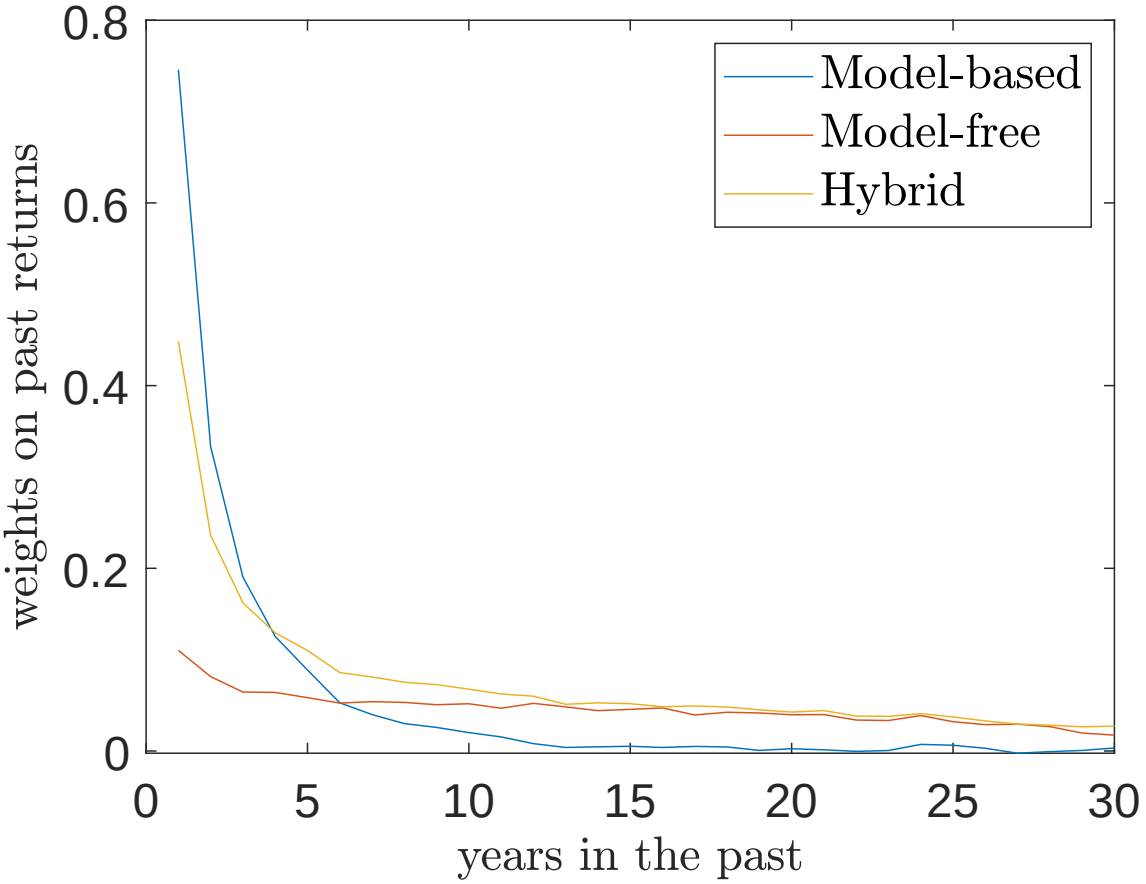

Figure A2. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors were exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system. There are 300,000 investors. Each investor believes there are two states, h and l. We set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu_h = 0.03$, $\mu_l = -0.01$, $\sigma = 0.2$, $\chi = 0.25$, and $w = 0.5$.

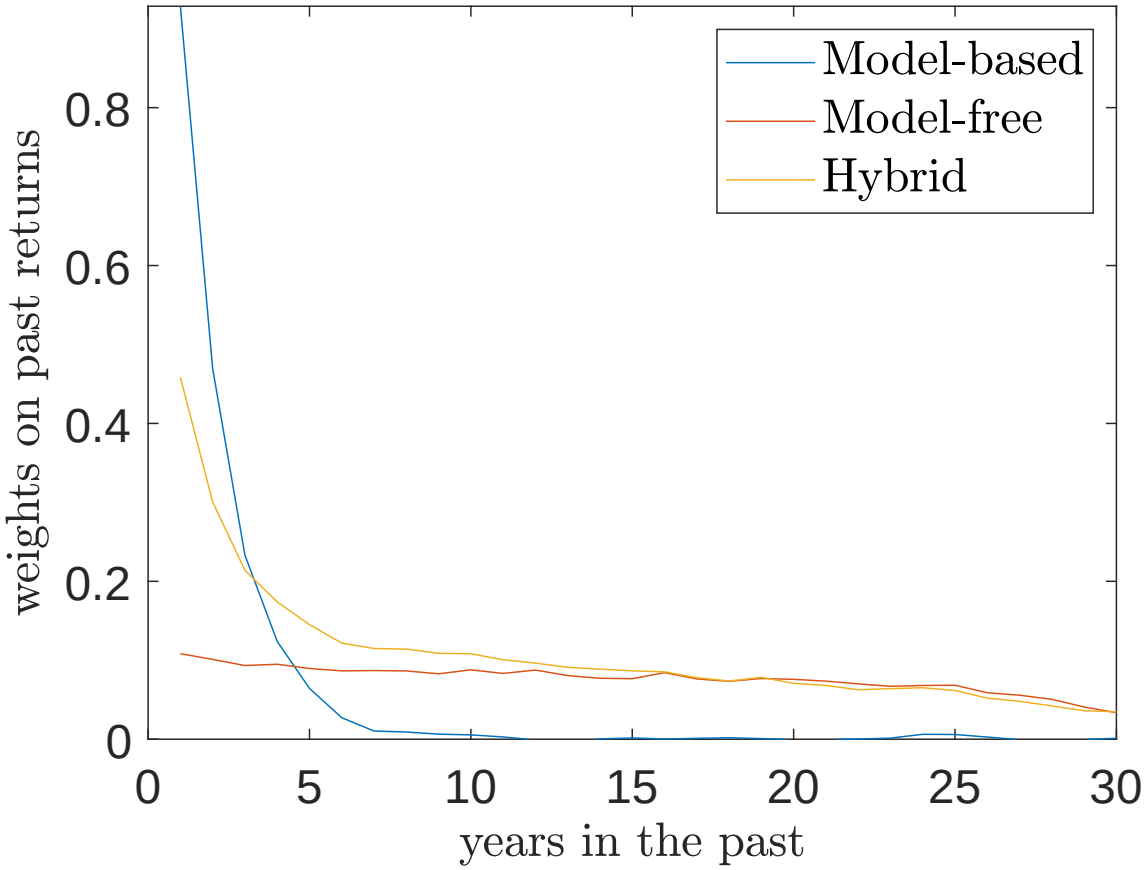

Figure A3. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors were exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system. There are 300,000 investors. Each investor believes there is one state. We set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu_h = 0.03$, $\mu_l = -0.01$, $\sigma = 0.2$, $\chi = 0.25$, and $w = 0.5$.