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

Nicholas Barberis and Lawrence Jin

July 2023*

Abstract

Motivated by neural evidence on the brain's computations, cognitive scientists are increasingly adopting a framework that combines two systems, namely "model-free" and "model-based" learning. We import this framework into a financial setting, study its properties, and use it to account for a range of facts about investor behavior. These include extrapolative demand, experience effects, the disconnect between investor allocations and beliefs in the frequency domain and the cross-section, the inertia in investors' allocations, and stock market non-participation. Our results suggest that model-free learning plays a significant role in the behavior of some investors.

*The authors' affiliations are Yale University and Cornell University, respectively; our e-mails are nick.barberis@yale.edu and lawrence.jin@cornell.edu. We are grateful to Andrew Caplin, Alex Chinco, Cary Frydman, Xavier Gabaix, Alex Imas, Chen Lian, Stefan Nagel, Elise Payzan-LeNestour, Indira Puri, Antonio Rangel, Josh Schwartzstein, Andrei Shleifer, Michael Woodford, and seminar participants at Caltech, Columbia University, Cornell University, Harvard University, Imperial College London, the University of California Los Angeles, the University of Chicago, the University of Pennsylvania, Washington University, Yale University, the AFA Annual Meeting and the NBER Behavioral Finance conference for very useful feedback. We are also grateful to Colin Camerer, Nathaniel Daw, Peter Dayan, Sam Gershman, John O'Doherty, and members of their lab groups for very helpful discussions about the psychological concepts in the paper. Steven Ma provided excellent research assistance.

1 Introduction

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

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

The model-based algorithm, by contrast, is similar to traditional economic approaches in that it does construct a model of the world – a probability distribution of future outcomes – and then uses this to compute the value of different actions. There are a number of model-based approaches; we use one that is often adopted in research in psychology and that, like

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

the model-free system, has neuroscientific support. Under this approach, after observing an outcome at some moment in time, the model-based system increases the probability it assigns to that outcome and downweights the probabilities of other outcomes. To do the updating, it again uses a learning rate and a prediction error that measures how surprising a realized outcome is; as before, there is evidence that the brain computes such prediction errors.

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

In this paper, we import this framework into a financial setting, study its implications for investor behavior, and use it to account for a range of empirical facts. To our knowledge, this is the first time the framework has been applied, in a comprehensive way, in an economic domain outside the laboratory. We choose a simple setting: one 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. The two algorithms tackle the problem in different ways. The model-based system learns a distribution of stock market returns over time by observing realized returns and uses it to decide on an allocation. The model-free system, by contrast, simply tries an allocation and observes the resulting portfolio return; if this return is good, the model-free system raises its estimate of the value of this allocation and is more likely to recommend this allocation again in the future.

We begin by characterizing investor behavior in our framework, paying particular attention to the model-free system – for economists, the more novel system. Specifically, we look at how the stock market allocation proposed by each of the model-free and model-based systems depends on past stock market returns. The model-based allocation puts weights on past market returns that are positive and that decline for more distant past returns. We find that the model-free system also recommends an allocation that puts positive weight on past returns, and show that it does so through a mechanism that is new to financial economics. In brief: A good stock market return reinforces the investor's previous allocation, whether the allocation was low or high. However, this reinforcement is stronger when the prior allocation is high: for a given positive market return, the reward, or portfolio return, is higher when the

prior allocation is high. As a consequence, on average, a good stock market return leads the investor to subsequently take a higher allocation.

We also find that, relative to the model-based system, whose recommended allocation puts heavy weight on recent returns, the model-free allocation puts substantially more weight on distant past returns. This is because it updates slowly: since it learns from experience, at each time, it updates only the value of the most recently-chosen allocation; the values of the other allocations are unchanged and hence depend only on more distant past returns. It therefore takes a long time for the influence of past returns to fade.

We then use our framework to shed light on some important facts about investor behavior, investor beliefs, and the relationship between the two. This is striking because, in prior research, this framework has been used primarily to explain behavior in experimental settings; it is notable, then, that it can also account for real-world financial behavior.

A prominent idea, motivated by empirical evidence, is that investors have extrapolative demand: their demand for a risky asset depends on a weighted average of the asset's past returns, where the weights are positive and larger for more recent returns. The analysis summarized above shows that model-free and model-based learning can both offer a foundation for extrapolative demand; the model-free system, in particular, does so in a way that is new to financial economics.

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

Individual investors overreact to recent market returns when forming beliefs about future market returns: as shown by Greenwood and Shleifer (2014) among others, their beliefs depend strongly on recent returns even though there is little autocorrelation in realized returns. Our framework captures this overreaction through the model-based system: after a good market return, this system increases the probability it assigns to good returns, leading the investor to expect a higher market return in the future.

Beyond simply capturing overreaction in beliefs, our framework can also resolve two puzzling disconnects between investors' beliefs and stock market allocations. While individual investor beliefs about future stock market returns depend primarily on recent past market returns, Malmendier and Nagel (2011) find that investors' allocations to the stock market depend significantly even on distant past market returns. We reconcile these findings by way of a deep property of our framework, which is that, of the two systems, only the model-based system has a role for beliefs: only this system explicitly constructs a probability distribution of future outcomes. When an individual is surveyed about his beliefs regarding future returns, he necessarily consults the model-based system – only this system can answer the survey question – and therefore gives an answer that depends primarily on recent past returns. However, his allocation is influenced by both the model-based and model-free systems and therefore depends significantly even on distant past returns.

Through a similar mechanism, our framework can also explain another disconnect between actions and beliefs, namely the low sensitivity of allocations to beliefs documented by several recent studies in the cross-section of investors.[2] If the stock market posts a high return, the investor's expectation about the future stock market return will go up significantly: the model-based system, which determines beliefs, puts substantial weight on recent returns. However, the investor's allocation will be less sensitive to the recent return: it is determined in part by the model-free system, which, relative to the model-based system, puts much less weight on recent returns.

Our framework can also help to account for some other empirical facts about investor behavior, 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. We also draw a number of predictions out of the framework – for example, that for an investor who is more confident in his beliefs, the brain is likely to assign more control to the model-based system, leading the investor's allocation to be more closely tied to his beliefs.

Since the model-free system learns slowly, it is not an efficient way of making investment decisions in real time. Nonetheless, for several reasons, it is likely, as our paper suggests, to influence financial decision-making. First, the model-free system is a fundamental component

[2]See Ameriks et al. (2020), Giglio et al. (2021), Charles, Frydman, and Kilic (2023), and Yang (2023).

of human decision-making. As such, it is likely to play a role in any decision unless explicitly "switched off" – and because it operates below the level of conscious awareness, many investors will not recognize its influence and will therefore fail to turn it off. Second, many people do not have a good "model" of financial markets – for example, they have a poor sense of the structure of asset returns. As a consequence, the brain is likely to assign at least some control of financial decision-making to the model-free system – again, without a person's conscious awareness – precisely because this system does not need a model of the environment. Third, for a less sophisticated investor, his model-free system may perform better than his model-based system, something that we demonstrate quantitatively. While the slow learning of the model-free system can be costly – this system is slow to learn genuinely useful information – it can also be beneficial, in that it leads the model-free system to exhibit only a mild form of the biased thinking embedded in the investor's model-based system.

Model-free learning algorithms are of interest not only to psychologists and neuroscientists, but also to computer scientists, albeit for a different purpose. Computer scientists see these algorithms as a powerful tool for solving challenging dynamic problems (Sutton and Barto, 2019). For example, these algorithms have been used 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 (Erev and Roth, 1998; Camerer, 2003, Ch. 6). One 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, albeit different from, the hybrid model we consider

below.

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

In Section 2, we formalize the model-free and model-based learning algorithms and show how they can be applied in a financial setting. In Section 3, we present an example to show how the two algorithms work and then study their implications for investor behavior. In Section 4, we use the framework to account for a range of facts about investor allocations and beliefs. Section 5 discusses some additional analysis while Section 6 concludes.

2 Model-free and Model-based Algorithms

To understand human decision-making, researchers in the fields of psychology and neuroscience are increasingly adopting a framework that combines model-free and model-based learning (Daw, Niv, and Dayan, 2005; Daw, 2014). In this section, we describe this framework and propose a way of applying it in a financial setting. Specifically, in Section 2.1, we describe the model-free algorithm; in Section 2.3, we lay out a model-based learning algorithm; and in Section 2.4, we show how the two algorithms are combined. In Section 2.2, 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. Along the way, we will also summarize some of the psychological and neuroscientific evidence for the framework.

2.1 Model-free learning

Model-free and model-based learning algorithms are intended to solve problems of the following form. Time is discrete and indexed by $t = 0, 1, 2, 3, \ldots$ At time t, the state of the world is denoted by s_t and an individual takes an action a_t. As a consequence of taking this action

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

in this state, the individual receives a reward r_{t+1} at time $t+1$ and arrives in state s_{t+1} at that time. The joint probability of s_{t+1} and r_{t+1} conditional on s_t and a_t is $p(s_{t+1}, r_{t+1}|s_t, a_t)$. The environment has a Markov structure: the probability of (s_{t+1}, r_{t+1}) depends only on s_t and a_t. In an infinite-horizon setting, the individual's goal is to maximize the expected sum of discounted rewards:

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

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

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

For "positive" applications, where we are trying to explain why people behave the way they do, the dynamic programming approach raises an obvious question. It may be hard to determine the probability distribution $p(\cdot)$; and even if we have a good sense of this distribution, it may be difficult, even for professional economists, to solve the Bellman equation for the value function. How, then, would an ordinary person be able to do so? Economists have long suggested that people act "as if" they have solved the Bellman equation – but they have not explained how this would come about. Psychologists, by contrast, have been trying to develop a more literal description of how people make decisions in dynamic settings – a framework that is rooted in the brain's actual computations. The leading such framework is the one we adopt in this paper, namely one that combines model-free and model-based learning.

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

to similar predictions.[4]

Q-learning works as follows. Let $Q^*(s,a)$ be the expected sum of discounted rewards – in other words, the value of the expression

$$E_t\left[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-(t+1)}r_\tau\right] \tag{2}$$

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

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

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t^{MF}[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s,a)], \tag{3}$$

where α_t^{MF} is known as the learning rate – the superscript stands for model-free – and where the term in square brackets is an important quantity known as the reward prediction error (RPE): the realized value of taking the action a – the immediate reward r_{t+1} plus a continuation value – relative to its previously anticipated value, $Q_t(s,a)$. Put simply, the updating rule in (3) says that, if, after taking the action a, the algorithm observes a better outcome than anticipated, it raises its estimate of the value of that action.

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

$$p(a_t = a|s_t = s) = \frac{\exp[\beta Q_t(s,a)]}{\sum_{a'}\exp[\beta Q_t(s,a')]}. \tag{4}$$

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

This probabilistic choice, known as a "softmax" specification, serves an important purpose: it encourages the algorithm to "explore," in other words, to try an action other than the one that currently has the highest Q value in order to learn more about the value of this other action. In the limit as $\beta \to \infty$, the algorithm chooses the action with the highest current Q value; in the limit as $\beta \to 0$, it chooses an action randomly. The parameter β is called the "inverse temperature" parameter, but we refer to it more simply as the exploration parameter. We discuss what exploration means in financial settings in more detail in Section 2.2.[5]

The algorithm is initialized at time 0 by setting $Q(s, a) = 0$ for all s and a. Consistent with (4), the time 0 action is chosen randomly from the set of possible actions. The process then proceeds according to equations (3) and (4). If the algorithm takes the action a in state s and this is followed by a good outcome, the value of $Q(s, a)$ goes up, making it more likely that, if the algorithm encounters state s again, it will again choose action a. Computer scientists have found Q-learning to be a useful way of solving the problem in (1); it can be shown that, under certain conditions, the Q values generated by the algorithm converge to the correct Q^* values (Watkins and Dayan, 1992).

Psychological background. While computer scientists make frequent use of model-free algorithms like Q-learning, what is more important for our purposes is that psychologists and neuroscientists are also interested in these algorithms. This is because of mounting evidence that they play an important role in human decision-making. This evidence comes in two forms: behavioral data – data on how people behave – and neural data on the brain's computations.

The behavioral data come from experimental paradigms that allow researchers to isolate the influence of model-free learning from more traditional model-based learning. One of the best known is the "two-step task" introduced by Daw et al. (2011). We summarize this task in Internet Appendix A. Analysis of participants' behavior in this experiment finds a large influence of model-free learning.[6]

Neural data has been an even bigger factor in the surge of interest in model-free learning.

[5]Another interpretation of the probabilistic choice in (4) is that it stems from cognitive noise: due to errors in perception or cognitive processing, the algorithm does not necessarily select the action with the highest Q value. See Woodford (2020) for a review of recent research on cognitive noise.

[6]Charness and Levin (2005) present a different experiment in which model-free and model-based learning – in their terminology, reinforcement learning and Bayesian learning – again make different predictions. They, too, find that participant behavior is guided to a significant extent by the model-free system. More recent experimental studies with a similar theme are Payzan-LeNestour and Bossaerts (2015) and Allos-Ferrer and Garagnani (2023).

A major finding in decision neuroscience is that the activity of certain neurons in the ventral striatum region of the brain lines up well with the reward prediction error used by model-free algorithms. This, in turn, suggests that the brain implements such model-free algorithms when making decisions. This observation was first made in influential papers by Montague, Dayan, and Sejnowksi (1996) and Schultz, Dayan, and Montague (1997). A large number of subsequent studies, using functional magnetic resonance imaging (fMRI) to study human decision-making, have presented similar neural evidence for model-free learning.[7]

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

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_{t,\pm}^{MF}(\text{RPE}), \tag{5}$$

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

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

2.2 A portfolio-choice setting

In Section 2.3, we lay out a model-based algorithm to complement the model-free algorithm of Section 2.1. Before we do so, it will be helpful to first describe the task that we 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 $R_f = 1$ in each period. The gross return on the risky asset

[7]See McClure, Berns, and Montague (2003), O'Doherty et al. (2003), Glascher et al. (2010), and Daw et al. (2011), among many others. Sutton and Barto (2019, Ch. 15) offer a useful review.

between time $t-1$ and t, $R_{m,t}$, where "m" stands for market, has a lognormal distribution

$$\begin{aligned} \log R_{m,t} &= \mu + \sigma\varepsilon_t \\ \varepsilon_t &\sim N(0,1), \text{ i.i.d.} \end{aligned} \tag{6}$$

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 2.1, so we use the notation a_t for it.[8] We construct an objective function that is realistic and also has the required form in (1). Specifically, the investor's goal is to maximize the expected log utility of wealth at some future horizon determined by his liquidity needs. The timing of these liquidity needs is uncertain; as such, the investor does not know in advance how far away this horizon is. More precisely, at time 0, the investor enters financial markets. If, coming into time $t \geq 1$, he is still present in financial markets, then, with probability $1-\gamma$, where $\gamma \in [0,1)$, a liquidity shock arrives. In that case, he exits financial markets and receives log utility from his wealth at time t. A short calculation shows that the investor's implied objective is then to solve

$$\max_{\{a_t\}} E_0 \left[\sum_{t=1}^{\infty} \gamma^{t-1} \log R_{p,t} \right], \tag{7}$$

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

$$R_{p,t} = (1-a_{t-1})R_f + a_{t-1}R_{m,t}. \tag{8}$$

Comparing (1) and (7), we see that this portfolio problem maps into the framework of Section 2.1: the generic reward r_t in equation (1) 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 (7). The solution is that, at each time t, the investor allocates the same constant fraction a^* of his wealth to the stock market, where

$$a^* = \arg\max_a E_t \log((1-a)R_f + aR_{m,t+1}). \tag{9}$$

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

The fact that the problem in (7) 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 (9) will *not* be intuitive, as 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 (7), then, as argued in the Introduction, there is reason to think that a model-free system like Q-learning will play a role in his decision-making. As a fundamental part of human thinking, the model-free system is likely to play a role in any decision unless it is explicitly turned off. And for an investor who is unsure about the distribution of asset returns, the brain is all the more likely to assign some control to the model-free system, precisely because this system does not rely on any information about this distribution. This leads to the question at the heart of this paper: How will an investor behave if model-free Q-learning influences his actions?

How can Q-learning be applied to the above problem? In principle, we could apply equation (3) 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).[9] While, for much of the paper, we put state dependence aside, we re-introduce it in Section 5 and summarize there an analysis in the Internet Appendix of the state-dependent case.

As in Section 2.1, then, let $Q^*(a)$ be the expected sum of discounted rewards – in other words, the value of

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

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

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

$$Q_{t+1}^{MF}(a) = Q_t^{MF}(a) + \alpha_{t,\pm}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)]. \tag{10}$$

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

$$p(a_t = a) = \frac{\exp[\beta Q_t^{MF}(a)]}{\sum_{a'} \exp[\beta Q_t^{MF}(a')]}. \tag{11}$$

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 (11) is central to the model-free algorithm and to the way psychologists think about human behavior. The term is less common in economics and finance. Nonetheless, many actions in financial settings can be thought of as forms of exploration – for example, any time an individual tries a strategy that is new to him, such as investing in a stock in a different industry or foreign country, or in an entirely new asset class. In our context, with one risk-free and one risky asset, exploration can be thought of as the investor choosing a different allocation to the stock market than before in order to learn more about the value of doing so.[10]

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

$$Q^*(a) = E\log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E\log((1-a^*)R_f + a^*R_{m,t+1}), \tag{12}$$

where a^* is defined in (9).

In the basic model-free algorithm in (10), after taking action $a_t = a$ at time t, only the Q

[10] As noted in Section 2.1, another possible foundation for the probabilistic choice in (11), one that may be relevant in financial settings, is cognitive noise.

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

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

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

where

$$\kappa(\widehat{a}) = \exp(-\frac{(\widehat{a}-a)^2}{2b^2}). \tag{14}$$

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 (14).[11]

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 (14), 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 $a + 10\%$.

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

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.[12]

We emphasize that the Q-learning algorithm above, with or without generalization, does not use any information about the distribution of risky asset returns in (6): 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. While the model-free system may appear uninformed, the fact that it uses so little information about the problem at hand is precisely what makes it powerful, in general: it can be applied in almost any setting. And despite being uninformed, its implications will turn out to be helpful for thinking about a range of facts in finance.

2.3 Model-based learning

Current research in psychology uses a framework in which decisions are guided by both model-free and model-based learning. Model-based systems, as their name indicates, build a model of the environment, which, more concretely, means a probability distribution of future outcomes – for example, in our setting, a probability distribution of stock market returns. There are various possible model-based systems. Which one should we choose? Our goal in this paper is to see if algorithms commonly used by psychologists can explain behavior in economic settings. We therefore take as our model-based system one that, like the model-free system of Section 2.1, is based on an algorithm that is used extensively by psychologists and is supported by neural evidence from decision-making experiments.

In our model-based system, an investor learns the distribution of stock market returns over time by observing realized market returns. At each date, he updates the probabilities of different returns using a prediction error analogous to the reward prediction error of Section

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

2.1. 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

$$p_{t+1}(R_m = R) = p_t(R_m = R) + \alpha_t^{MB}[1 - p_t(R_m = R)], \tag{15}$$

where α_t^{MB} is the model-based learning rate that applies from time t to time $t+1$. The term $1 - p_t(R_m = R)$ is a prediction error: the investor's prior estimate of the probability of the return equaling R was $p_t(R_m = R)$; when the return is realized, the probability of it equaling R is 1. After this update, the investor scales the probabilities of all other returns down by the same proportional factor so that the sum of all return probabilities continues to equal one. Since we are working with a continuous return distribution, we can assume that each return that is realized is one that has not been realized before. As such, $p_t(R_m = R) = 0$, which simplifies (15) 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: $R_{m,1}$, $R_{m,2}$, $R_{m,3}$, and $R_{m,4}$, at dates 1, 2, 3, and 4, respectively. The four rows below show the investor's perceived probability distribution of stock market returns at dates 1, 2, 3, and 4, in the case where the learning rate is constant over time, so that $\alpha_t^{MB} = \alpha$ for all t. In this notation, a comma separates a return from its perceived probability, while semicolons separate the different returns:

$$\begin{aligned}
&(R_{m,1}, 1) \\
&(R_{m,1}, 1-\alpha; R_{m,2}, \alpha) \\
&(R_{m,1}, (1-\alpha)^2; R_{m,2}, \alpha(1-\alpha); R_{m,3}, \alpha) \\
&(R_{m,1}, (1-\alpha)^3; R_{m,2}, \alpha(1-\alpha)^2; R_{m,3}, \alpha(1-\alpha); R_{m,4}, \alpha).
\end{aligned} \tag{16}$$

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 the reward prediction error used by model-free learning, so there is evidence that it encodes the prediction error used by model-

based learning.[13]

We noted in Section 2.1 that, when they implement model-free learning, psychologists allow for different model-free learning rates, α_+^{MF} and α_-^{MF}, for positive and negative reward prediction errors, respectively. We extend the model-based algorithm in a similar way, allowing for different model-based learning rates, α_+^{MB} and α_-^{MB}, for positive and negative net stock market returns, respectively. Specifically, following the gross return $R_{m,t+1} = R$,

$$p_{t+1}(R_m = R) = \alpha_{t,+}^{MB} \text{ for } R > 1, \tag{17}$$

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

$$p_{t+1}(R_m = R) = \alpha_{t,-}^{MB} \text{ for } R \leq 1, \tag{18}$$

with the probabilities of all other returns being scaled down by $1-\alpha_{t,-}^{MB}$. The different learning rates can be thought of as reflecting a different level of attention to, or a different level of concern about, positive as opposed to negative outcomes (Kuhnen, 2015).

With this perceived return distribution in hand, how does the investor come up with a model-based estimate of $Q^*(a)$, the value of choosing an allocation a on some date and then continuing optimally thereafter? We again follow an approach taken by experimental studies in decision neuroscience (Glascher at al., 2010). We assume that, for any allocation a, the individual computes his time t model-based estimate of $Q^*(a)$, denoted $Q_t^{MB}(a)$, by taking the correct form of $Q^*(a)$ in equation (12) and applying it for his *perceived* time t return distribution:

$$Q_t^{MB}(a) = E_t^p \log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E_t^p \log((1-a_t^*)R_f + a_t^* R_{m,t+1}), \tag{19}$$

where

$$a_t^* = \arg\max_a E_t^p \log((1-a)R_f + aR_{m,t+1}) \tag{20}$$

and where (19) differs from (12) only in that the expectation E under the correct distribution

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

has been replaced by the expectation E_t^p under the investor's perceived distribution at time t.

While our financial setting is a simple one, it is rich enough to create a tension between the model-free and model-based systems. If the investor starts with a low allocation to the stock market and the market then posts a high return, the model-free system wants to stick with a low allocation because this action was "reinforced": it was followed by a positive reward prediction error. In intuitive terms, since the investor's action is "working," there is no need to change it. 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.[14]

The model-free and model-based systems are not the only learning algorithms the brain uses. Another important class of algorithms are "observational learning" algorithms which learn by observing the actions and outcomes of other people (Charpentier and O'Doherty, 2018). There is also some evidence for "counterfactual learning" algorithms which learn about the value of actions not taken. We focus on the model-free and model-based algorithms because they have received the most attention from cognitive scientists; because they likely "span" other algorithms, in that these other learning systems tend to generate predictions that lie somewhere between those of the model-free and model-based systems; and because these other algorithms are not necessary for our purpose: as we show in Section 4, a simple combination of model-free and model-based learning alone can account for several aspects of investor behavior.

2.4 A hybrid framework

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

$$Q_t^{HYB}(a) = (1-w)Q_t^{MF}(a) + wQ_t^{MB}(a), \tag{21}$$

[14] As described in Internet Appendix A, a similar tension is present in experimental studies of model-free and model-based learning.

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:

$$p(a_t = a) = \frac{\exp[\beta Q_t^{HYB}(a)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(a')]}. \tag{22}$$

In this paper, we focus on the case where w is constant over time, as this already leads to a rich set of properties and applications. Nonetheless, a well-known hypothesis in psychology is that w varies over time: at each moment, the brain puts more weight on the system it deems more "reliable" at that point (Daw, Niv, and Dayan, 2005). In Section B of the Internet Appendix, we formalize and explore this idea using an implementation proposed by researchers in decision neuroscience in which a system's reliability is measured by the absolute magnitude of its prediction errors: if the model-free reward prediction errors have been large in absolute magnitude, the brain deems the model-free system to be less reliable and raises w, thereby allocating more control to the model-based system. We discuss the implications of this idea in Section 5 and in the Internet Appendix.[15]

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, $\{R_{m,-L+1}, \ldots, R_{m,0}\}$; uses these to form a perceived distribution of market returns as in (17) and (18); and then computes model-based Q values

[15]The model-free and model-based learning framework is not without critics. For example, Feher da Silva et al. (2023) question a subset of the evidence for the framework. However, they do not offer a concrete alternative, and the model-free and model-based learning framework continues to be the leading approach to thinking about a large body of both behavioral and neural evidence.

by way of that distribution, as in (19).[16]

3 Properties of Investor Behavior

We begin this section with an example that illustrates the mechanics of the model-free and model-based systems. We then study the implications of the framework for investor behavior. 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 this analysis to account for several facts about investor behavior.

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

The four learning rates – α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} – play an important role in our framework. How should they be set? If we were taking a normative perspective – if we wanted to use the algorithms of Section 2 to solve the problem in (7) as efficiently as possible – the answer would be to use learning rates that decline over time. Specifically, the time t

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

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

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

model-based learning rates in (17) and (18) would be

$$\alpha_{t,+}^{MB} = \alpha_{t,-}^{MB} = 1/(L+t+1), \tag{23}$$

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

$$\sum_{t=0}^{\infty} \alpha_{t,\pm}^{MF} 1_{\{a_t=a\}} = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} (\alpha_{t,\pm}^{MF})^2 1_{\{a_t=a\}} < \infty, \tag{24}$$

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

In this paper, however, we are taking a "positive" perspective – our goal is to explain observed behavior. What matters for our purposes is therefore not the learning rates people should use, but rather the learning rates they actually use. Psychology research does not offer definitive guidance on people's learning rates, but most studies of actual decision-making use learning rates that are constant over time (Glascher et al., 2010); moreover, the recorded activity of neurons that encode the reward prediction error is consistent with a constant learning rate (Bayer and Glimcher, 2005). For this reason, we focus on constant learning rates. To start, we give all investors the same constant learning rates. Later, we allow for dispersion in these rates across investors.[19]

3.1 An example

To show how the model-free and model-based systems work, we start with an example. Throughout the paper, we use the same baseline parameter values, in part for consistency and in part to show that a fixed set of parameter values can account for a range of observed facts. We consider an investor who is exposed to a sequence of stock market returns from $t = -L$ to $t = T$, where $L = T = 30$. The returns are simulated from the distribution in (6) 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

[19] One reason that has been proposed for why the human learning system would use constant learning rates is that these are well-suited to the non-stationary environments that humans often encountered during the evolutionary process.

parameter β to 30, the discount factor γ to 0.97 – this corresponds to an expected investment horizon of 33 years – and the degree of generalization b to 0.0577.[20] At each date, we allow the investor to choose his stock market allocation a_t from one of 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$. We later examine how our results depend on the values of all the key parameters.

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 (11), (13), and (14) (upper panel) and the model-based Q values, Q^{MB}, based on equations (19) and (20) (lower panel) that the investor assigns to the 11 allocation strategies on his first six dates of participation in financial markets, namely $t = 0$, 1, 2, 3, 4, and 5. The rows labeled "net market return" show the net return of the stock market at each date. In each column, the number in bold corresponds to the action that was taken in the previous period; for example, the number -0.065 in bold at date 1 in the upper table indicates that the investor chose a 70% allocation at date 0.[21]

Consider the upper panel of Table 1. The model-free system begins operating at time 0. At that time, then, it assigns a Q value of zero to all the allocations. It then randomly selects the allocation 70%. The net stock market return at time 1 is negative, which means that the investor's net portfolio return and reward prediction error are also negative. The time-1 Q value for the 70% allocation therefore falls below zero. As per equations (13) and (14), 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.

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

[21] In the case where decisions are determined by the model-based system alone, we assume that the investor still chooses actions probabilistically, in a manner analogous to that in (11). In our setting, for the model-based system, this probabilistic choice does not offer the usual exploration benefits: in each period, the investor learns the same thing about the distribution of stock market returns regardless of which allocation he chooses. We keep the probabilistic choice to allow for a more direct comparison with the model-free system – but also because, if, as suggested earlier, this stochastic choice stems in part from cognitive noise, it will be relevant for model-based learning too. For these reasons, whenever we consider the model-based system in isolation, we will allow for probabilistic choice.

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

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

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

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.

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

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

3.2 Dependence on past market returns

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

To study this, we take $300,000$ investors and expose each of them to a different sequence of simulated stock market returns from $t = -L$ to $t = T$. We then take investors' stock market allocations a_T at time T, regress them on the past 30 annual stock market returns

$\{R_{m,T},\ R_{m,T-1},\ldots,R_{m,T-29}\}$ the investors have been exposed to, and record the coefficients. We do this for three cases, namely those where investor allocations are determined by the model-free system alone; by the model-based system alone; and by the hybrid system. For all investors, as before, we set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. For ease of interpretation, we turn off generalization for now, so that $b = 0$.[22] Finally, we set $w = 0.5$, so that the hybrid system puts equal weight on the model-free and model-based systems. We later look at how changing the values of key model parameters affects the results.[23]

Figure 1 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 more gradual for the model-free system, a property that will play a key role in some of our applications. Given that the hybrid system combines the model-free and model-based systems, it is natural that the line for the hybrid system is, approximately, a mix of the model-free and model-based lines.

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

[22] We use "$b = 0$" as shorthand for model-free learning without generalization. When $b = 0$, we compute model-free Q values using equation (10) rather than equations (13)-(14), although the latter equations give the same result as $b \to 0$.

[23] The goal function in (7) 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.

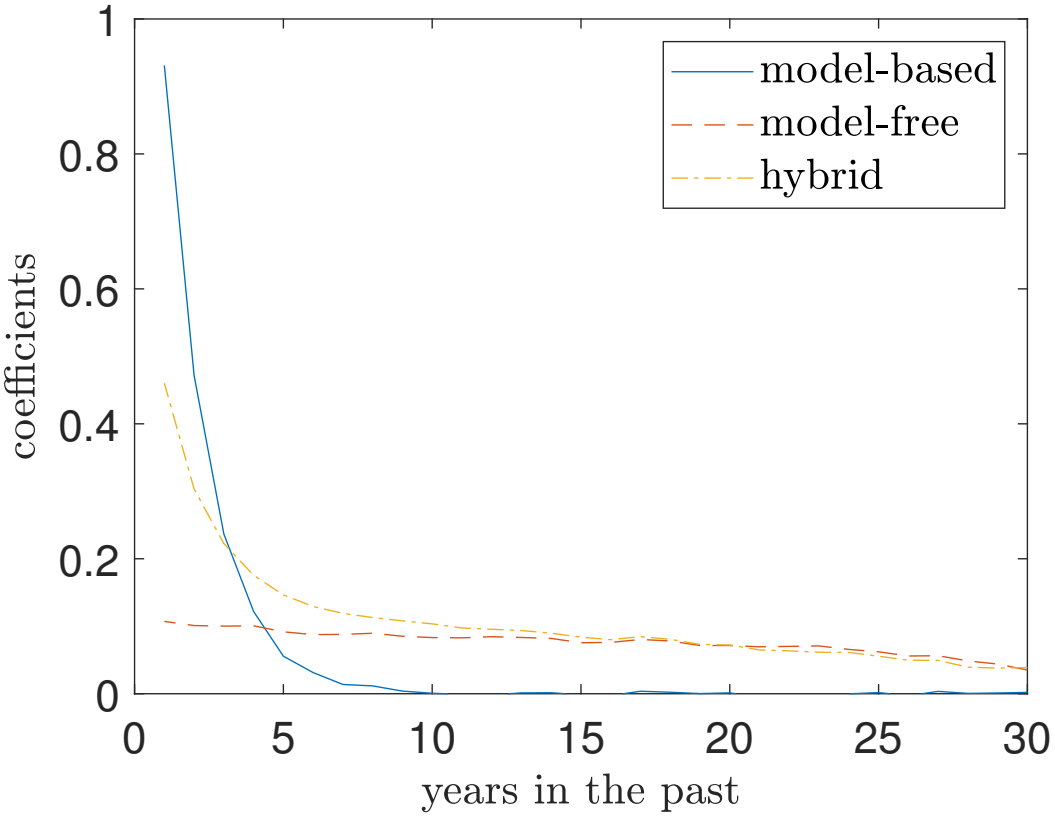

Figure 1. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system. The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

The intuition in the case of the model-free system is different and, to our knowledge, new to financial economics. If the investor chooses a 20% stock market allocation and the market posts a high return, this "reinforces" the action of choosing a 20% allocation: the positive reward prediction error raises the Q value of this allocation, making it more likely that the investor will choose it again in the future. Similarly, if he chooses an 80% allocation and the market posts a high return, this reinforces the 80% allocation. In one case, then, a high market return leads the investor to choose a low allocation; in the other, it leads him to choose a high allocation. Why then, on average, does a high market return lead to a higher allocation, as shown by the dashed line in Figure 1? The reason is that the reinforcement is stronger in

the case of the 80% allocation: a high stock market return leads to a larger reward prediction error when the investor's prior allocation is 80% than when it is 20%. As such, the net effect of a good stock market return, after averaging over the possible prior allocations, is to lead the investor to choose a high stock market allocation.

We now explain why the weights that the two systems put on past market returns decline as we go further into the past. In the case of the model-based system, this is because, when this system updates its perceived return distribution after seeing a new stock market return, it scales down the probabilities of earlier returns, reducing their importance. Intuitively, by using a constant learning rate, the investor is acting as if the environment is non-stationary; as such, he puts greater weight on recent returns. The top graph in Figure 2 shows how the time T allocation recommended by the model-based system depends on past stock market returns for four different values of the learning rates α_+^{MB} and α_-^{MB}, namely 0.05, 0.1, 0.2, and 0.5. The graph shows that, regardless of the learning rate, the allocation puts weights on past returns that are positive and that decline the further back we go into the past, with the decline being more pronounced for higher learning rates.

Figure 1 shows that, for the model-free system, the weights on past returns again decline as we go further into the past, but much more gradually. Why is this? When the model-free system updates the Q value of an action, this tends to downweight the influence of past returns on this Q value, relative to the most recent return. However, this effect passes through to allocation choice in a much more gradual way than for the model-based system because, at each time, the model-free system primarily updates only one Q value, namely that of the most recently-chosen action; as such it takes much longer for past returns to lose their influence on the investor's allocation.[24] The bottom graph in Figure 2, which plots the relationship between the model-free allocation and past returns for four different values of the learning rates α_+^{MF} and α_-^{MF}, shows that the model-free allocation typically puts positive and declining weights on past returns, with the decline being more pronounced for higher learning rates.

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

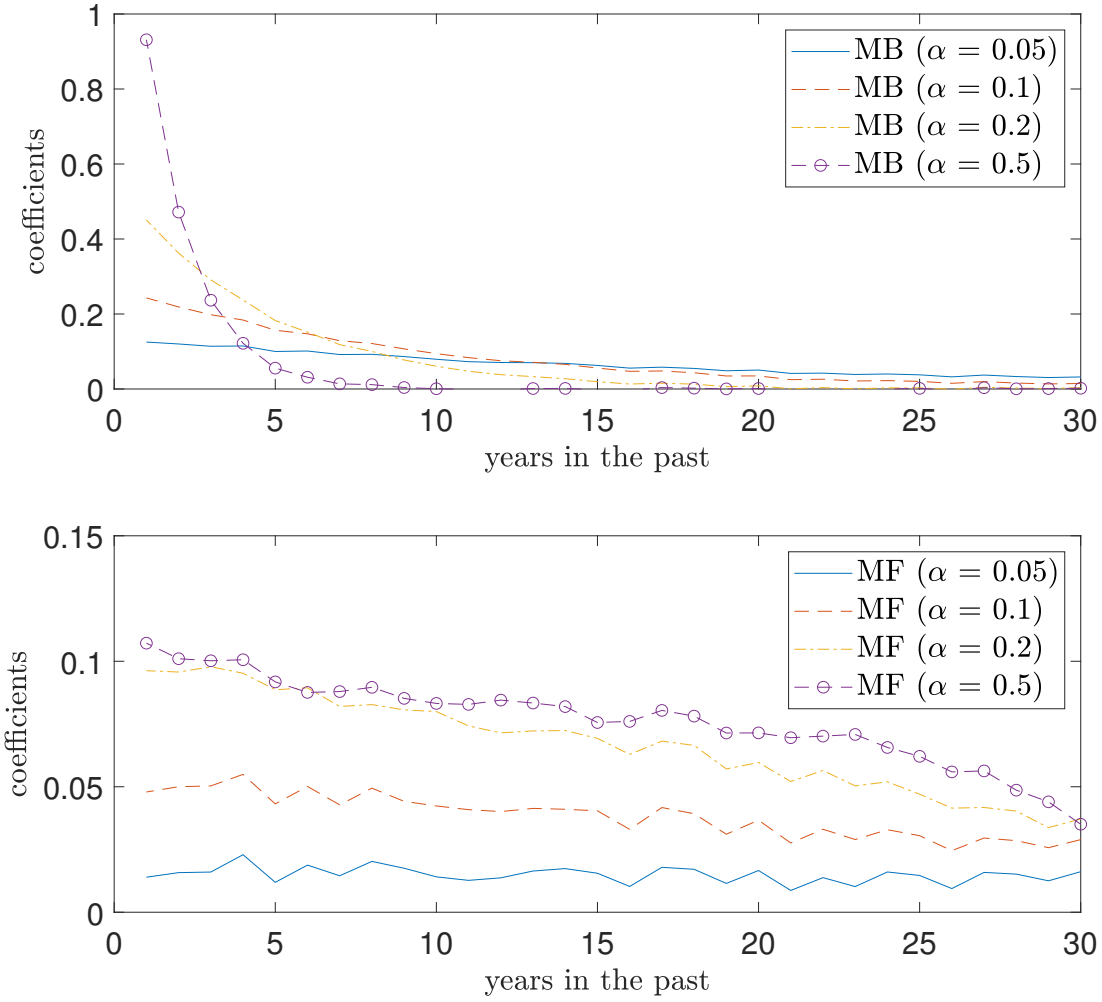

Figure 2. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to. The top graph plots the coefficients for the model-based system for four values of the learning rates α_+^{MB} and α_-^{MB}, namely 0.05, 0.1, 0.2, and 0.5. The point on the horizontal axis that marks j years in the past corresponds to the coefficient on $R_{m,T+1-j}$. The bottom graph plots the coefficients for the model-free system for four values of the learning rates α_+^{MF} and α_-^{MF}, namely 0.05, 0.1, 0.2, and 0.5. There are 300,000 investors. We set $L = T = 30$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

The graphs in Figure 3 show how the relationship between investors' time T model-free allocations and past stock market returns changes as we vary one of the model parameters

while keeping the others at their benchmark levels. Across the four graphs, we vary the degree of generalization, the degree of exploration, the discount factor, and the number of allocation choices. Changing these parameters would have little effect on model-*based* allocations. However, Figure 3 shows that it has significant impact on model-free allocations. While for many parameter values, including those used in Figures 1 and 2, the model-free allocation puts more weight on recent than on distant past returns, Figure 3 shows that, for some parameter values, it can put more weight on distant than on recent past returns. Moreover, the figure shows the conditions under which this happens – for example, for higher degrees of generalization. We explain the full intuition for the patterns in Figure 3 in Internet Appendix C.[25]

While the model-free algorithm is simple to state – it is summarized in equation (13) – it is difficult to derive analytical results about its predictions. Nonetheless, for certain special cases, we *are* able to derive such results, which are precisely about the dependence of model-free allocations on past market returns and which, to our knowledge, are the first results of their kind. We present these results and their proofs in Theorems 1 to 4 of Internet Appendix D. Specifically, we show that, under some conditions, as $t \to \infty$, the sensitivity of the expected allocation at time t to the market return k periods earlier, $R_{m,t-k} = R$, is given by

$$\frac{\partial E(a_t)}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^3}\left(\frac{R^\beta+1-\alpha R^\beta}{R^\beta+1}\right)^k \tag{25}$$

for the model-free allocation, and by

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

for the model-based allocation, where α is the constant learning rate for both systems. These results are consistent with the patterns in Figure 1. Expressions (25) and (26) both decline monotonically as k increases. Moreover, the model-free coefficient in (25) is lower than the model-based coefficient in (26) for low values of k, but higher than the model-based coefficient for high values of k. This provides an analytical foundation for the property we have

[25]The results in Figures 1 to 3 are for an annual-frequency implementation of our framework. We have studied the effect of changing the frequency. If we fix the learning rates $\alpha_{\pm}^{MB}$ and $\alpha_{\pm}^{MF}$ but switch to a semi-annual, quarterly, or monthly implementation, this has a significant effect on the model-based allocation – it depends all the more on recent returns – but a much smaller impact on the model-free allocation. As such, implementing the framework at a higher frequency creates a larger wedge between the two systems.

emphasized in this section, namely that, relative to the model-based system, the model-free system puts significantly more weight on distant past returns.

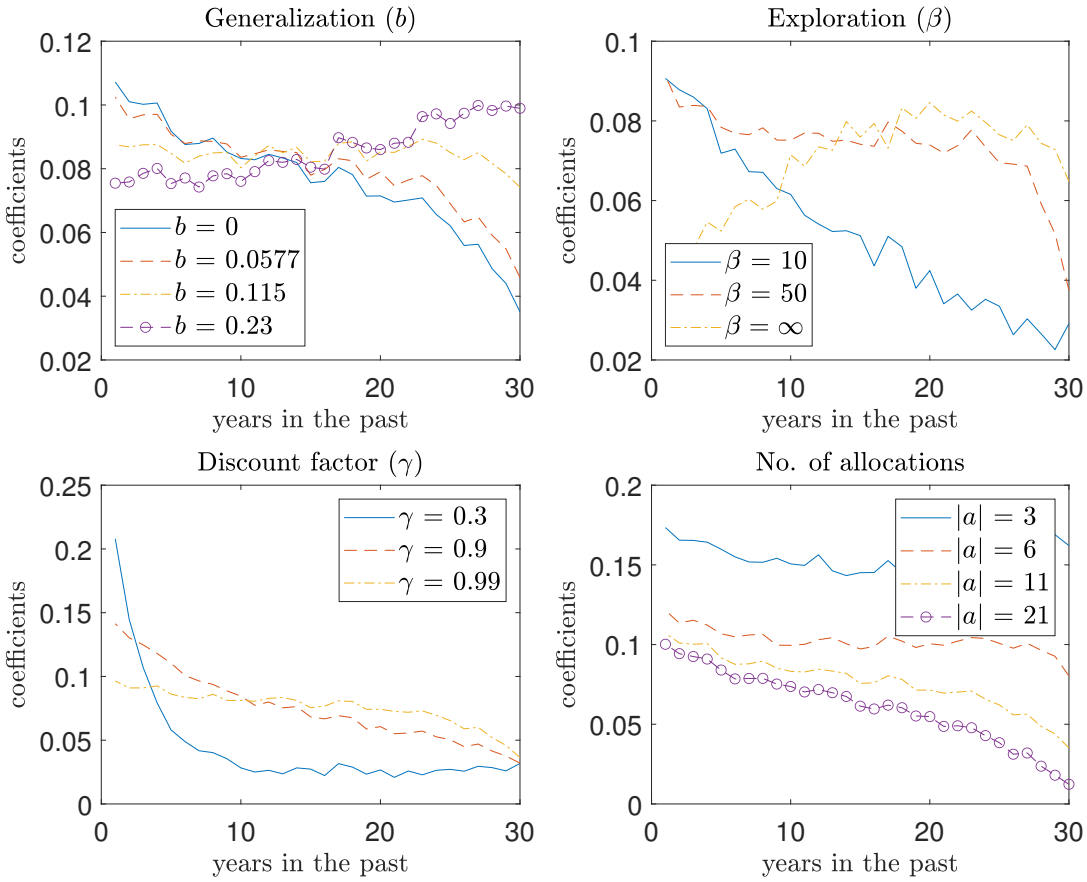

Figure 3. For different sets of parameter values, we run a regression of investors' allocations to the stock market a_T at time T under the model-free system on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients. The lines in the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

3.3 Dependence on past allocations and portfolio returns

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

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

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

$$a_T = 0.2 + 0.63a_{T-1} - 0.35r_{p,T} + 0.75a_{T-1}r_{p,T} + \varepsilon_T. \tag{27}$$

The relationship in (27) captures four features of the model-free system, which are most easily illustrated with numerical examples.[26]

First, equation (27) shows that the model-free allocation at time T is closely tied to the previous period's allocation. For example, if $a_{T-1} = 20\%$ and the net portfolio return is a

[26]Here, and in some other places in the paper, we study the implications of behavior when it is determined by the model-free system alone. We do this in order to understand the model-free system more deeply. However, the prevailing view in psychology is that behavior is driven by both the model-free and model-based systems, in combination. As such, it does not make sense to compare (27) directly to observed behavior. Rather, it is the combined impact of model-free and model-based learning that should be compared to observed behavior, as it will be in Section 4.

neutral $r_{p,T} = 0$, then the expected time T allocation $E(a_T) = 33\%$; and if $a_{T-1} = 80\%$ and $r_{p,T} = 0$, then $E(a_T) = 71\%$. In both cases, the expected time T allocation is fairly close to the time $T-1$ allocation. This is because, at each time, the model-free system primarily updates the Q value of the most recently-chosen action. As such, the Q values at time T are similar to the Q values at time $T-1$; this, in turn, means that the time T allocation is likely to resemble the time $T-1$ allocation.

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

Third, regression (27) captures the reinforcing effect of the portfolio return. If $a_{T-1} = 20\%$ and the portfolio return is a neutral $r_{p,T} = 0$, then $E(a_T) = 33\dot{3}\%$; but if $r_{p,T} = 0.2$, then $E(a_T) = 29\%$: the high portfolio return reinforces the 20% allocation and pulls the time T allocation towards it, from 33% down to 29%. Similarly, if $a_{T-1} = 80\%$ and $r_{p,T} = 0$, then $E(a_T) = 71\%$; but if $r_{p,T} = 0.2$, then $E(a_T) = 76\%$: this time, the high portfolio return reinforces the 80% allocation and pulls the time T allocation towards it, from 71% up to 76%.

Finally, regression (27) captures the indirect way that the *market* return affects the model-free allocation. If $a_{T-1} = 20\%$ and the net market return is a neutral $r_{m,t} \equiv R_{m,t} - 1 = 0$, then $r_{p,T} = 0$ and $E(a_T) = 33\%$ as before. But if $r_{m,T} = 0.2$, then $r_{p,T} = 0.04$ and $E(a_T) = 32\%$. In this case, the high market return modestly reinforces the prior allocation of 20% and lowers the expected allocation by 1%, from 33% to 32%. Similarly, if $a_{T-1} = 80\%$ and $r_{m,T} = 0$, then $r_{p,T} = 0$ and $E(a_T) = 71\%$. But if $r_{m,T} = 0.2$, then $r_{p,T} = 0.16$ and $E(a_T) = 75\%$. In this case, the high market return strongly reinforces the prior allocation of 80% and increases the expected allocation by 4%, from 71% to 75%. Averaging across the two prior allocations, the positive market return increases the investor's time T allocation by approximately $(4-1)/2 = 1.5\%$. This provides a numerical illustration of the mechanism described in the previous section: the model-free allocation depends positively on past market

returns because a high market return generates greater reinforcement when the investor's prior allocation is high.[27]

3.4 Variability and dispersion

Regression (27) in the previous section shows that model-free allocations are "sticky": the allocation at any time hews closely to the allocation in the previous period. This suggests that the model-free system will lead to less variability in an investor's allocation over time. We now document this more formally.

To demonstrate the result, we first allow for dispersion in learning rates across investors.[28] For each investor, we draw each of their learning rates – each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} – from a uniform distribution centered at $\bar{\alpha}$ and with width Δ. As before, the parameter values are $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$; and as in Section 3.1, we have $b = 0.0577$, so that there is some generalization. We set the new parameter Δ to 0.5. We take 1,000 investors, expose them to the same sequence of stock market returns from $t = -L$ to $t = T$, and compute the variability in their allocations: for each investor in turn, we compute the standard deviation of his allocations $\{a_{T-j}\}_{j=0}^{30}$ over time and then average these standard deviations across investors. We repeat this exercise 300 times for different return sequences and average the resulting variability measures.

The solid and dashed lines in the top three graphs in Figure 4 plot the variability of investor allocations under the model-based and model-free systems, respectively, as we vary the values of three parameters – the exploration parameter β, the mean learning rate $\bar{\alpha}$, and the dispersion Δ of learning rates – while keeping the other parameter values fixed at their benchmark levels. The graphs confirm that the model-free system leads to lower variability than the model-based system: the dashed lines are substantially below the solid lines.

[27] A more precise calculation, which averages across all eleven possible prior allocations, leads to a similar result. The 1.5% number approximately matches the prediction of the dashed line in Figure 1 for the sensitivity of the model-free allocation to a 20% net market return, namely $(0.1072)(0.2) = 2.14\%$, where 0.1072 is the coefficient on the most recent market return in a regression of model-free allocations on past market returns.

[28] 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.

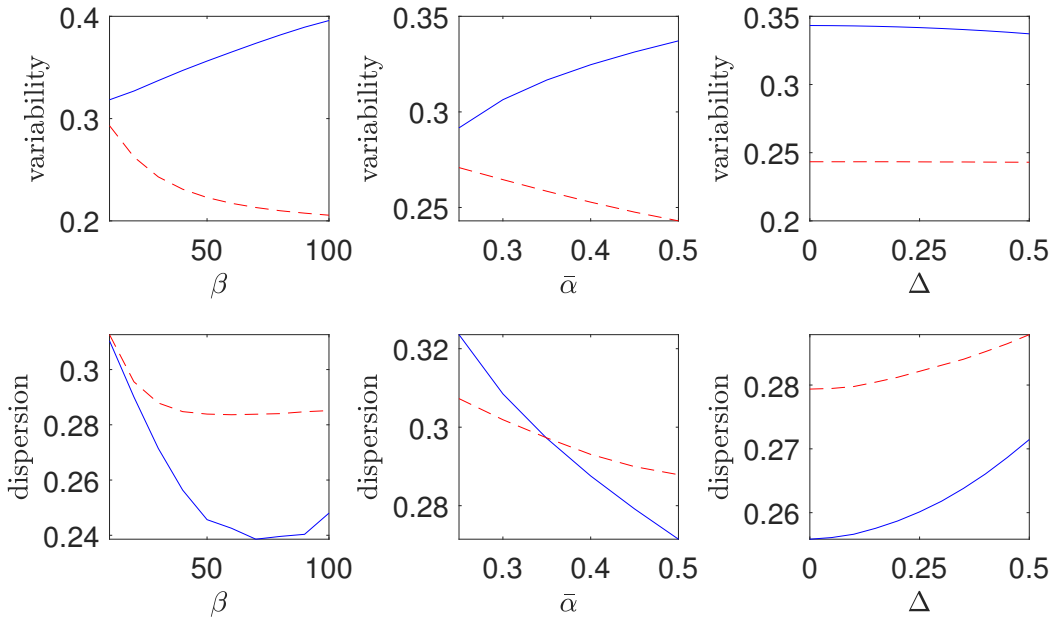

Figure 4. The upper graphs plot the variability of stock market allocations – the standard deviation of allocations between time 0 and time T, computed for each investor in turn and averaged across investors. The lower graphs plot the dispersion, across investors, of their stock market allocations at time T. The solid and dashed lines correspond to the model-based and model-free systems, respectively. For each system, the graphs vary the exploration parameter β, the mean learning rate $\bar{\alpha}$, or the dispersion in learning rates Δ, while keeping the other parameter values fixed at their benchmark levels. The results are averaged across 300 simulations; each simulation features 1,000 investors, all of whom see the same return sequence. The benchmark parameter values are $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

Using the simulated data from the above exercise, we also compute another quantity that will be helpful when we turn to applications, namely the dispersion in allocations across investors at time T. For each of our 300 simulations, and for each of the model-free and model-based systems, we compute the standard deviation of the 1,000 investors' time T allocations and then average these estimates across the 300 simulations. The lower graphs in Figure 4 plot the resulting dispersion measures as we vary each of β, $\bar{\alpha}$, and Δ. The graphs show that the two systems generate similarly high dispersion in investor allocations. We return to this finding in Section 4.

4 Applications

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

We have associated the risky asset in our framework with the aggregate stock market. Our applications therefore focus on important facts about this market – facts about investor allocations, investor beliefs, and the relationship between the two. We address these facts by way of a simple parameterization of our framework, where, by "simple," we mean that each investor's learning rates α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} are constant over time, and, for all investors, the values of these learning rates are drawn from the same distribution. Our initial goal is not to provide a close quantitative fit to observed facts; it is to show that a simple parameterization can provide a qualitative, and approximate quantitative, fit to the data. Toward the end of this section, we estimate the parameter values that provide a closer quantitative match to the data.

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

Given the above structure, at time T, the cross-section of investors resembles the one we

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

Each investor in the economy is trying to solve the problem in (7) and chooses allocations from the set {0%, 10%,. . . , 90%, 100%} according to the hybrid system in (21)-(22). 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 this section in order to show that a single parameterization is consistent with a range of empirical facts. As in Section 3, we set $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$, so that investors put equal weight on the model-free and model-based systems. Later, we formally estimate the value of w that best fits the data.

4.1 Extrapolative demand

Our first application builds on the analysis of Section 3.2. A common assumption in psychology-based models of asset prices and investor behavior is that people have extrapolative demand: their demand for a financial asset depends positively on the asset's past returns, and especially on its recent past returns.[29]

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

[29]A partial list of papers that study extrapolative demand, either theoretically or empirically, is Cutler, Poterba, and Summers (1990), De Long et al. (1990), Barberis and Shleifer (2003), Barberis et al. (2015, 2018), Cassella and Gulen (2018), Bastianello and Fontanier (2022), Chen, Liang, and Shi (2022), Jin and Sui (2022), Liao, Peng, and Zhu (2022), and Pan, Su, Wang, and Yu (2023).

than if he had a low allocation; a high allocation therefore receives more reinforcement, making it more likely that he will choose a high allocation going forward.

To confirm that the framework of Section 2 generates extrapolative demand, we run a regression of investors' allocations a_T at time T, as determined by the hybrid system, on the past stock market returns each of them has observed. The relationship between the allocation and past returns is plotted as the solid line in Figure 5. The graph confirms that an investor's allocation to the stock market is a positive function of the market's past returns, with weights that decline the further back we go into the past. The solid line in Figure 5 is similar to the line marked "Hybrid" in Figure 1 in that both lines correspond to decisions made under the hybrid system. However, the two lines differ because, relative to the analysis in Section 3.2, we are now allowing for dispersion across investors in their learning rates and for multiple cohorts. The multiple cohorts in particular make the solid line in Figure 5 decline more quickly than the "Hybrid" line in Figure 1: some of the investors in the market at time $T = 30$ entered only at time 25; as such, their model-free system puts no weight on returns before time 25.

Recent models of extrapolative demand have taken it to be a consequence of investor beliefs. Our analysis shows that it can also result from a mechanism that is unrelated to beliefs and based instead on the reinforcement of past actions. More precisely, our framework says that extrapolative demand has two sources: a model-based source derived from beliefs that puts heavy weight on recent returns, and a model-free source that puts substantial weight even on distant past returns. We exploit this two-component structure of extrapolative demand in Sections 4.3 and 4.4 to shed light on some puzzling disconnects between allocations and beliefs.

Prior work has shown that certain types of extrapolative demand can help explain the excess volatility in stock markets and the predictability of market returns (De Long et al., 1990; Barberis et al., 2015). Given that the investors in our framework have a form of extrapolative demand, it is natural to expect that, in an asset pricing setting, these investors will generate excess volatility and predictability. This excess volatility will be driven in part by beliefs – this is the influence of the model-based system – but also by a new mechanism, namely model-free reinforcement of past actions. For space reasons, we do not present a formal analysis of asset prices in the current paper, but such an analysis is feasible and a natural direction for future research.

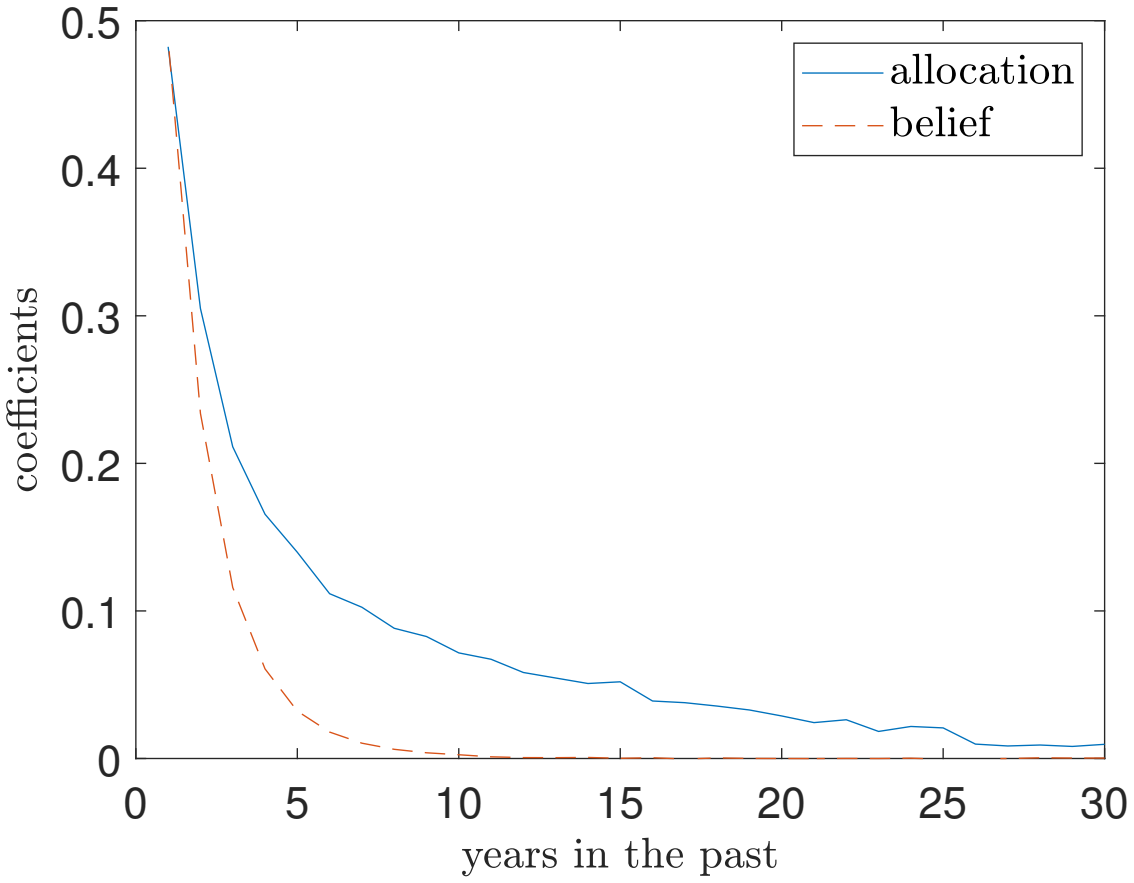

Figure 5. The solid line plots the coefficients in a regression of the stock market allocation a_T at date T chosen by investors who use a hybrid system to make decisions on the past 30 years of stock market returns the investors were exposed to. The dashed line plots the coefficients in a regression 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 $w = 0.5$.

4.2 Experience effects

Malmendier and Nagel (2011) show that investors' decisions are affected by their experience: whether an investor participates in the stock market, and how much he allocates to the stock market if he does participate, can be explained in part by the stock market returns he has

personally experienced – in particular, by a weighted average of the returns he has personally lived through, with more weight on more recent returns.

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

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

$$(n+1-j)^{\lambda}/A, \qquad j=1,2,\ldots,n, \tag{28}$$

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

To see if our framework can generate these two features of experience effects, we proceed as follows. For each of the six cohorts, we take the $50,000$ investors in the cohort and regress their time T allocations a_T on the past 30 years of stock market returns. Figure 6 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.

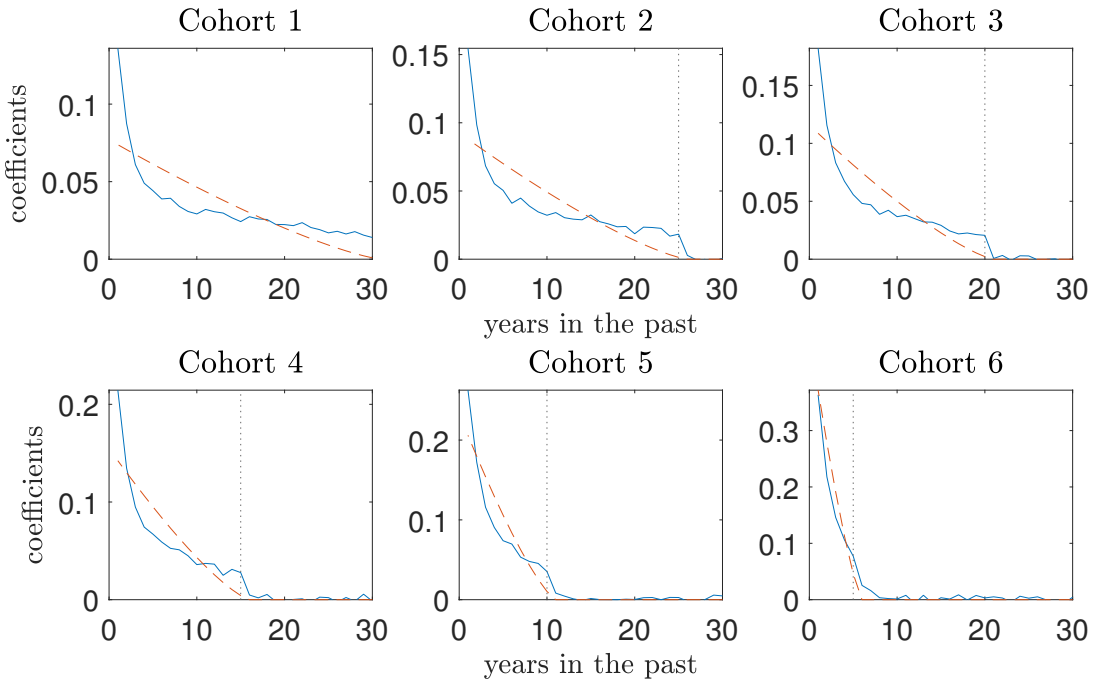

Figure 6. The six graphs correspond to six cohorts of investors. In each graph, the solid line plots the coefficients, normalized to sum to one, in a regression of the time T stock market allocations a_T of the investors in that cohort on the past 30 years of stock market returns they were exposed to. The six cohorts have different numbers of years of experience, namely $n = 5$, 10, 15, 20, 25, and 30; the vertical dotted line in each graph marks the time at which the cohort enters financial markets. There are 300,000 investors, with 50,000 in each cohort. For each investor, each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} is drawn independently from a uniform distribution with mean $\bar{\alpha}$ and width Δ. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$. In each graph, the dashed line plots a functional form for experience effects calibrated to data by Malmendier and Nagel (2011), namely $(n + 1 - j)^\lambda / A$, where j is the number of years in the past, $\lambda = 1.3$, and A is a normalizing constant.

By comparing the solid and dashed lines for each graph in turn, we see that our framework can capture both aspects of experience effects. Consider the bottom-left graph for cohort 4 which enters at date 15. The solid line shows that our framework generates a kink in the dependence of allocation on past market returns as we move from a return these investors experienced – the return 15 years in the past – to one they did not experience, the return 16 years in the past. The kink is driven by investors' model-free system, which puts substantial

weight even on a return experienced 15 years in the past, but no weight at all on returns before that. The graph also shows that, within the subset of returns that these investors experience, their allocation puts greater weight on more recent past returns. Both the model-free and model-based systems contribute to this pattern, although the model-based system does so more.

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

Using an analogous approach, 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).[30]

4.3 Investor beliefs: Overreaction and the frequency disconnect

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

Aside from capturing overreaction in beliefs, our framework also resolves two puzzling disconnects between investor beliefs and investor actions – one in the frequency domain, which we discuss in this section, and one in the cross-section of investors, which we address in the

[30]See Malmendier and Wachter (2022) for a review of other psychology-based foundations for experience effects.

next section. We account for these puzzles by way of a deep property of our framework, namely that, of the two systems, only the model-based system has an explicit role for beliefs. The model-free system, by contrast, has no notion of beliefs: it does not construct a probability distribution of future outcomes; instead, it learns the value of actions simply by trying them and observing the outcomes.

The disconnect in the frequency domain is simple to state. While investor expectations about future returns depend heavily on *recent* past returns, investor stock market allocations depend to a substantial extent even on distant past returns (Malmendier and Nagel, 2011).[31]

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

Figure 5 illustrates these points. As discussed in Section 4.1, the solid line shows how *allocations* depend on past returns: it plots the coefficients in a regression of investors' allocations to the stock market at time T on the past 30 years of stock market returns they were exposed to. The dashed line shows how *beliefs* depend on past returns: it plots the coefficients in a regression of investors' expectations at time T about the future one-year stock market return on the past 30 years of stock market returns they were exposed to. Comparing the two lines, we see that, while beliefs depend primarily on recent returns, allocations depend significantly even on distant past returns.

A number of studies find a positive time-series correlation between investor beliefs and allocations. For example, Greenwood and Shleifer (2014) find that the average investor ex-

[31]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 $\lambda \approx 1.3$. 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 $\lambda \approx 37$, which puts much more weight on recent returns.

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

Our results also have implications for our understanding of asset prices. Recent research has tried to explain excess volatility in the stock market with models in which investors form beliefs about future market returns by extrapolating past returns (Barberis et al., 2015; Jin and Sui, 2022). One challenge faced by these models is that, when calibrated to survey data, they predict too low a persistence for the price-dividend ratio: since beliefs depend primarily on recent returns, they have low persistence, and prices inherit this property. Our framework offers a solution to this problem. In a market with investors of the type studied here, beliefs, generated by the model-based system, will have low persistence because they load heavily on recent returns. However, prices are determined by both the model-free and model-based systems, and will therefore have high persistence: since the model-free system loads even on distant past returns, it generates a high-persistence component in allocations, and hence in prices too.

4.4 Investor beliefs: The cross-sectional disconnect

Using survey responses from Vanguard investors, as well as data on these investors' allocations to the stock market, Giglio et al. (2021) document another disconnect between beliefs and actions. Regressing investors' stock market allocations on investors' expected one-year stock market returns, they obtain a coefficient approximately equal to one. By contrast, a traditional Merton model of portfolio choice predicts a much higher coefficient. A similar insensitivity of allocations to beliefs is also documented, using a variety of approaches, by Ameriks et al. (2020), Charles, Frydman, and Kilic (2023), and Yang (2023).

Our framework can help explain this insensitivity. The mechanism is similar to that for the frequency disconnect: it again relies on the fact that, while an investor's allocation is based on both the model-free and model-based systems, only the model-based system has an explicit role for beliefs. To see the implications of this, suppose that the stock market posts a

high return. The investor's expectation about the future stock market return will then go up significantly: the model-based system, which determines beliefs, puts substantial weight on recent returns. However, the investor's allocation will be less sensitive to the recent return: it is determined in part by the model-free system, which, relative to the model-based system, puts much less weight on recent returns.

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

In Section 2.4, we noted that, according to a well-known hypothesis in psychology, the brain puts more weight on the learning system it views as more reliable. This idea – one that we analyze formally in Internet Appendix B – leads to one of the clearest predictions of our framework, namely that, if an individual is more confident in his beliefs, his actions will be more sensitive to his beliefs. Confidence in one's beliefs is a sign that the model-based system – the system that generates beliefs – is more reliable. The brain therefore allocates more control to it, leading to a higher sensitivity of allocations to beliefs. Giglio et al. (2021) offer evidence consistent with this: they find that, for the subsample of people who say they are more confident in their beliefs, allocations are indeed more sensitive to beliefs.

4.5 Dispersion and inertia

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

The lower panel of Figure 4 shows that the model-based system generates substantial dis-

persion in allocations. This dispersion is primarily due to differences in beliefs across investors, which in turn are driven by differences in learning rates; the probabilistic choice further adds to the dispersion. The lower panel shows that the model-free system also generates substantial dispersion in allocations. This is striking because this dispersion cannot be easily attributed to differences in beliefs or objective functions: as noted earlier, the model-free system has no notion of beliefs; moreover, in our setting, all investors have the same objective function in (7). Instead, the differences in allocations recommended by the model-free system are due to the process of decision-making itself. The probabilistic choice leads investors to try different allocations in their early years of financial market participation. Different allocations are then reinforced for different investors, which leads to differences in allocations even many years later.

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

The framework in this paper offers a new way of thinking about inertia in investor holdings: it says that the inertia arises endogenously from the model-free system. Regression (27) in Section 3.3 shows that an investor's model-free allocation in any period is closely tied to his allocation in the previous period. More directly, the upper panel of Figure 4 shows that, relative to the model-based system, the model-free system generates lower variability, or equivalently, higher inertia. The reason is that the model-free system learns slowly: at each time, it primarily updates only the value of the most recently-chosen allocation. This, in turn, increases the likelihood that the allocation at time t will be similar to the allocation at time $t-1$.

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

which they will respond to an uninformed demand shock.

4.6 Non-participation

A long-standing question asks why many U.S. households do not participate in the stock market despite its substantial risk premium. Our framework can shed light on this. In particular, the model-free system tilts investors toward not participating. To see why, consider an investor who makes decisions according to the model-free system. If he allocates some money to the stock market but then experiences a poor market return, this lowers the Q values of the chosen allocation and of those similar to it; this, in turn, raises the probability that, in a subsequent period, he will switch to a 0% allocation to the market. Importantly, if he does move to a 0% allocation, the model-free system will stop learning about the stock market: realized market returns will no longer affect the reward prediction error, and so, even allowing for generalization, these returns will not be reflected in the updated Q values. As such, the investor fails to learn that the stock market has better properties than indicated by the one poor return he experienced. This will tend to keep him at a 0% allocation for an extended period of time.

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

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

The results confirm that the model-free system tilts investors toward non-participation. We find that, under the model-free system, 43% of investors are not participating – in other

words, are at a 0% allocation – for at least 80% of the 30 dates from $t = 1$ to $t = 30$. By contrast, under the model-based system, fewer than 1% of investors spend more than 80% of the time not participating. In a similar vein, under the model-free system, 59% of investors have a non-participation streak that is at least 10 years long; under the model-based system, only 7% of investors have a streak of this length. The simulated data also support the mechanism for non-participation laid out above. We find that, under the model-free system, long streaks of non-participation are typically preceded by a poor experienced stock market return. The longer the non-participation streak, the more negative the prior experienced return, on average.

The mechanism we have used to think about non-participation can also be deployed to explain, more generally, why households make persistent investment mistakes. Suppose that, when trying to achieve some financial goal, an investor has ten options to choose from – ten different financial products, say, or ten different investment strategies. In this context, as in others, the model-free system will be slow to learn: at each time, it learns only about the option it is currently trying. As a consequence, an investor who is influenced by model-free learning will take a long time to figure out which of the options is the best one.

4.7 Parameter estimation

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

We explain the estimation procedure in full in Internet Appendix E and summarize it here. We do the estimation in two steps. We first use data on investor beliefs to estimate $\bar{\alpha}^{MB}$. We then estimate $\bar{\alpha}^{MF}$, β, and w by targeting the allocation-belief sensitivity and the experience

effect. We keep the remaining parameters at their benchmark values from before, namely $L = T = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.[32]

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

With this value of $\bar{\alpha}^{MB}$ in hand, we move to the second step: we search for values of $\bar{\alpha}^{MF}$, β, and w that best match two empirical targets. The first is the coefficient in a regression of investor allocations on investor beliefs, which Giglio et al. (2021) find to be approximately one. For given values of $\bar{\alpha}^{MF}$, β, and w, we can compute this coefficient in simulated data from our framework. Our second target is the functional form in (28) with $\lambda = 1.3$, which Malmendier and Nagel (2011) use to capture the relationship between allocations and past returns. Intuitively, we are looking for values of $\bar{\alpha}^{MF}$, β, and w that minimize the distance between unnormalized versions of the solid and the dashed lines in the six graphs in Figure 6.

We find that the parameter values that best match the allocation-belief sensitivity and the experience effect are $\bar{\alpha}^{MF} = 0.26$, $\beta = 20$, and $w = 0.38$. In words, to match the data, our framework requires substantial weight on both the model-free and model-based systems. The reason is the following. As shown by the dashed lines in Figure 6, the experience effect we are trying to capture involves both a steep initial decline in the coefficients on past returns, but also a significant dependence on distant past experienced returns. The upper panel of Figure 2 shows that the model-based system can capture the steep initial decline in coefficients, but, when calibrated to do so, it cannot capture the dependence on distant past returns. By contrast, the lower panel of Figure 2 shows that the model-free system can capture a high dependence on distant past returns but not the initial decline. To match both parts of the experience effect, we need to put substantial weight on both systems. The significant weight on the model-free system also helps to explain the low sensitivity of allocations to beliefs.

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

5 Additional Analysis

In this section, we discuss some additional analysis. We start by comparing the performance of the model-free and model-based systems. We then turn to some extensions of the framework.

5.1 Performance of the two systems

We have noted two reasons why model-free learning may play at least some role in investor decision-making: it is likely to engage automatically whenever the investor is experiencing rewards; and for an investor who feels that he does not have a good model of the environment, the brain is all the more likely to assign some control to the model-free system.

There is one more reason why the model-free system may influence decision-making. For an investor with a poor understanding of financial markets, and whose model-based system is therefore flawed, the model-free system's performance may be at least as good as that of the model-based system. As a consequence, even if the investor becomes aware of the influence of the model-free system on his behavior, he may continue to rely on it.

We can illustrate this point quantitatively. For the setting of Section 2 with i.i.d market returns, and for the parameter values and simulation structure in the caption to Figure 1, we find that the performance of the model-free system is similar to that of the model-based system. When investors use only the model-free system to make decisions, the mean and standard deviation of their per-period excess portfolio returns between $t = 0$ and $t = 30$, averaged across the 300,000 investors in the simulation, are 1.72% and 12.69%, respectively. By comparison, for investors who use only the model-based system, the corresponding numbers are 1.58% and 12.66%.

These numbers may understate the effectiveness of the model-free system. In Internet Appendix F, we analyze the case where the i.i.d return structure of Section 2 is replaced with one that captures the long-run mean-reversion seen in some asset classes. In words, if a weighted average of the risky asset's prior returns is high, then its subsequent mean return is low; if its prior returns are moderate, then its subsequent mean return is also moderate; and if its prior returns are low, then its subsequent mean return is high. In this case, we find that the state-independent model-free algorithm of Section 2 outperforms the state-independent model-based algorithm of that section by a substantial margin: it generates a mean and standard

deviation of portfolio returns equal to 1.61% and 12.99%, respectively; the corresponding numbers for the model-based system are 0.96% and 12.77%.

There is a good reason why, for less sophisticated investors, their model-free systems may outperform their model-based systems. The model-free system learns slowly. This has an obvious cost: the model-free system is slow to learn genuinely useful information. However, it also has a benefit: it leads the model-free system to exhibit only a mild form of the biased thinking built into some investors' model-based systems. For example, as shown in Section 3, the model-free system is less extrapolative, and this is valuable in a market with mean-reverting returns.

5.2 Extensions of the framework

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

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

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

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

State dependence. Thus far, our learning algorithms have not allowed for state dependence: we have worked with action values $Q(a)$ rather than state-action values $Q(s, a)$ because even this simple case has many applications. In Internet Appendix F, we examine the predictions of our framework when we allow for state dependence. In particular, we consider the setting with state-dependent returns summarized in Section 5.1 and study the implications for investor behavior when both the model-free and model-based algorithms take the state dependence into account. Echoing a result reported above, we find that the model-free and model-based systems have fairly similar performance: when decisions are made by the model-free system, the mean and standard deviation of excess portfolio returns are 1.77% and 12.98%, respectively. When decisions are made by the model-based system, the corresponding numbers are 1.94% and 13.25%. Once again, the slow learning of the model-free system has both costs and benefits: this system is slower to detect the mean-reversion in asset returns; however, it also exhibits a less extrapolative asset demand, which is valuable when returns are mean-reverting.

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

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

Alternative action spaces. Throughout the paper, we have used a standard action

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

Inferring beliefs from the model-free system. Until now, we have associated beliefs only with the model-based system. However, it is possible that an individual may also use the model-*free* system to make inferences about beliefs. For example, when an investor is asked for his beliefs about the stock market's future return or risk, it is natural that he will consult the model-based system, which will give him a direct measure of beliefs. However, he may also be influenced by the model-free system, and if $Q^{MF}(a = 1) > Q^{MF}(a = 0)$, so that his model-free system assigns the stock market a higher Q value than the risk-free asset, he may take this as a sign that the stock market has better *properties*, on several dimensions – for example, both a higher expected return and lower risk. This can help explain Giglio et al.'s (2021) finding that, when investors expect high returns in the stock market, they simultaneously expect the market to have lower risk, contrary to the prediction of traditional frameworks in which subjectively perceived risk and return are positively related.[33]

6 Conclusion

When economists try to explain human decision-making in dynamic settings, they typically assume that people are acting "as if" they have solved a dynamic programming problem. By contrast, cognitive scientists are increasingly embracing a different approach, one based on model-free and model-based learning. In this paper, we import this framework into a simple financial setting, study its implications for investor behavior, and use it to account for a range of facts about investor allocations, investor beliefs, and the relationship between the two. Through the model-based system, our framework preserves a role for beliefs in driving

[33]By way of a similar mechanism, our framework can also generate Hartzmark, Hirshman, and Imas' (2021) finding that ownership amplifies the extent of extrapolation in investor beliefs.

investor behavior. However, through the model-free system, it also introduces a new way of thinking about this behavior, one based on reinforcement of past actions.

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

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

7 References

Afrouzi, H., Kwon, S., Landier, A., Ma, Y., and D. Thesmar (2023), "Overreaction in Expectations: Theory and Evidence," *Quarterly Journal of Economics* 138, 1713-1764.

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

Allos-Ferrer, C. and M. Garagnani (2023), "Part-time Bayesians: Incentives and Behavioral Heterogeneity in Belief Updating," *Management Science*, forthcoming.

Ameriks, J., Kezdi, G., Lee, M., and M. Shapiro (2020), "Heterogeneity in Expectations, Risk Tolerance, and Household Stock Shares: The Attenuation Puzzle," *Journal of Business and Economic Statistics*, 38, 633-646.

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

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

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

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

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

Bastianello, F. and P. Fontanier (2022), "Expectations and Learning from Prices," Working paper.

Bayer, H. and P. Glimcher (2005), "Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal," *Neuron* 47, 129-141.

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

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

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

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

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

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

Charles, C., Frydman, C., and M. Kilic (2023), "Insensitive Investors," Working paper.

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

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

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

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

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

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

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

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

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

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

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

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

Feher da Silva, C., Lomardi, G., Edelson, M., and T. Hare (2023), "Rethinking Model-based and Model-free Influences on Mental Effort and Striatal Prediction Errors," *Nature Human Behavior* 7, 956-969.

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

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

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

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

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

Hartzmark, S., Hirshman, S., and A. Imas (2021), "Ownership, Learning, and Beliefs," *Quarterly Journal of Economics* 136, 1665-1717.

Huang, X. (2019), "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 (2022), "Asset Pricing with Return Extrapolation," *Journal of Financial Economics* 145, 273-295.

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

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

Kuhnen, C. (2015), "Asymmetric Learning from Financial Information," *Journal of Finance* 70, 2029-2062.

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

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

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

Malmendier, U. and J. Wachter (2022), "Memories of Past Experiences and Economic Decisions," Working paper.

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

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

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

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

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

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

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

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

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

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

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

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

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

Yang, J. (2023), "On the Decision Relevance of Subjective Beliefs," Working paper.

INTERNET APPENDIX

A. Experimental Evidence of Model-free Behavior

A number of experimental paradigms allow researchers to isolate the influence of model-free learning from model-based learning. Among the best known is the "two-step task" introduced by Daw et al. (2011). In this section, we summarize this task and some key findings about behavior in the task.

In the first stage of the experiment – see Figure A1 – a participant is given a choice between two options, A and B. If he chooses A, then, with probability 0.7, he is given a choice between options C and D, and with probability 0.3, a choice between options E and F. Conversely, if he chooses B in the first stage, then, with probability 0.7, he is given a choice between E and F, and with probability 0.3, a choice between C and D. After choosing between C and D or between E and F, the participant receives the reward associated with the chosen second-stage option. He repeats this task multiple times with the goal of maximizing the sum of his rewards.[1]

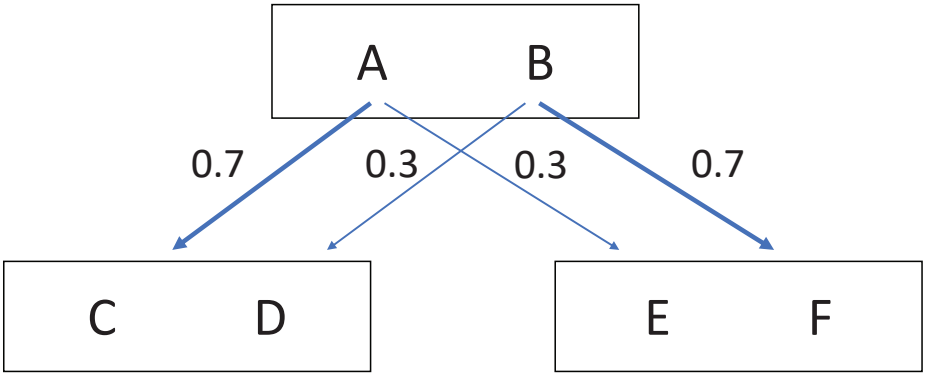

Figure A1. The diagram shows the structure of an experiment in Daw et al. (2011). In the first stage, the participant has a choice between two options, A and B; in the second stage, he chooses either between options C and D or between options E and F. The arrows indicate the transition probabilities from the first to the second stage. After making a choice at the second stage, the participant receives the reward associated with the chosen option.

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

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

structure of the task; since B offers a greater likelihood of ending up with the rewarded option E, he prefers B.

To evaluate the relative influence of model-free and model-based thinking on people's choices, Daw et al. (2011) run a regression of whether a participant repeats his previous first-stage choice on two variables: an indicator variable that equals one if this previous choice resulted in a reward; and this indicator interacted with another indicator variable that equals one if the individual saw the common rather than the rare second-stage options. For example, following an initial choice of A, the common second-stage options are C and D while the rare ones are E and F. If behavior is driven purely by the model-free system, only the coefficient on the first regressor will be significant. If behavior is driven purely by the model-based system, only the coefficient on the second regressor will be significant. The authors find that both coefficients are significant, which means that both systems are playing a role; an estimation exercise indicates that participants are putting approximately 60% weight on the model-free system and 40% on the model-based system.[2]

The above experiment illustrates a tension between the model-free and model-based systems. To repeat: if an individual chooses A and then E and is rewarded, the model-free system wants to repeat action A in the next round, while the model-based system, recognizing that B is more likely to lead to E, wants to choose B. The same tension is present in the financial market setting we lay out in Section 2.2 of the paper. If the investor starts with a low allocation to the stock market and the market then posts a high return, the model-free system wants to stick with a low allocation because this action was reinforced: it was followed by a positive reward prediction error. By contrast, the model-based system wants to increase the investor's allocation to the stock market: it now perceives a more attractive distribution of market returns and wants more exposure to it.

The presence of both model-free and model-based influences on behavior is also supported by neural data. We discuss some of this evidence in Section 2.1 of the main text.

B. Time-varying Weights on the Two Systems

In the main text, we take the weight on the model-based system to be constant over time. A well-known hypothesis in psychology is that this weight varies over time: at each moment, the brain puts more weight on the system it deems more reliable at that time (Daw, Niv, and Dayan, 2005). We now formalize this idea and examine its implications in our setting. To do this, we borrow a specification from neuroscience in which the reliability of a system is measured by the absolute magnitude of its past prediction errors (Lee, Shimojo, and O'Doherty, 2014): if a system's prediction errors are lower in absolute magnitude, it is deemed more reliable and given more weight in decision-making.[3]

Let w_{MB} be the weight assigned to the model-based system and $w_{MF} = 1 - w_{MB}$ the

[2]Feher da Silva et al. (2023) suggest that behavior in the two-step task may be driven by switching between different model-based systems, rather than by a combination of model-free and model-based learning. However, they do not offer a concrete alternative to the model-free and model-based learning approach, and the latter continues to be the dominant framework for thinking about a large body of both behavioral and neural data.

[3]For more details, see pages 19 to 22 of the Supplementary Material for Lee, Shimojo, and O'Doherty (2014).

weight assigned to the model-free system. The dynamics of w_{MB} are given by

$$\frac{dw_{MB}}{dt} = \alpha_t(1 - w_{MB}) - \beta_t w_{MB}, \tag{1}$$

where α_t is the transition rate from the model-free system to the model-based system and β_t is the transition rate from the model-based system to the model-free system. Since we have a discrete-time setting, we replace (1) by

$$w_{MB,t+1} - w_{MB,t} = \alpha_{t+1}(1 - w_{MB,t}) - \beta_{t+1} w_{MB,t}, \tag{2}$$

where the weight w_{MB} is bounded between zero and one.

The transition rates are functions of system reliability:

$$\alpha(\chi_{MF}) = \frac{A_\alpha}{1 + \exp(B_\alpha \cdot \chi_{MF})}, \quad \beta(\chi_{MB}) = \frac{A_\beta}{1 + \exp(B_\beta \cdot \chi_{MB})}, \tag{3}$$

where χ_{MF} and χ_{MB} are measures of the reliability of the model-free and model-based systems, respectively; A_α and A_β are the maximum transition rates; and B_α and B_β are the sensitivities of the transition rates to the reliability measures.

We construct the reliability measure χ_{MF} for the model-free system following Lee, Shimojo, and O'Doherty (2014). We classify past reward prediction errors (RPEs) into three categories: positive RPEs, negative RPEs, and RPEs that are close to zero. Specifically, if $\text{RPE} > \delta$, this RPE is classified as a positive RPE; if $\text{RPE} < -\delta$, it is a negative RPE; and if $-\delta \leq \text{RPE} \leq \delta$, it is a near-zero RPE. Suppose that model-free learning begins operating at time 0, when all the model-free Q values are zero, and that the current time is t. We count the total number of positive RPEs, the total number of negative RPEs, and the total number of near-zero RPEs:

$$\#\text{RPE}_+ = \sum\nolimits_{i=1}^{t} \mathbf{1}_{\text{RPE}_i > \delta}, \quad \#\text{RPE}_- = \sum\nolimits_{i=1}^{t} \mathbf{1}_{\text{RPE}_i < -\delta}, \quad \#\text{RPE}_0 = \sum\nolimits_{i=1}^{t} \mathbf{1}_{-\delta \leq \text{RPE}_i \leq \delta}. \tag{4}$$

The reliability measure χ_{MF} at time t is then defined as

$$\chi_{MF} = \frac{\chi_{MF,0}}{\chi_{MF,+} + \chi_{MF,-} + \chi_{MF,0}}, \tag{5}$$

where

$$\chi_{MF,i} = \frac{(3 + \sum_j \#\text{RPE}_j)(4 + \sum_j \#\text{RPE}_j)}{(2 + \sum_{j \neq i} \#\text{RPE}_j)}, \quad i \in \{+, -, 0\}. \tag{6}$$

The intuition of the above measure is that the model-free system is deemed more reliable if a larger fraction of its past prediction errors are close to zero: it is straightforward to show that χ_{MF} is increasing in the fraction of time an investor experiences near-zero RPEs.

We now turn to χ_{MB}, the reliability measure for the model-based system. In our setting, the prediction error for this system is always equal to one. As such, the reliability measure in (5), modified for the model-based system, implies a constant χ_{MB}. We therefore set β_t to a constant: $\beta_t = \beta(\chi_{MB}) \equiv \beta_{MB}$.

To examine the properties of the above framework, we need to set the values of four

parameters: δ, A_α, B_α, and β_{MB}. We set δ equal to the sample standard deviation of RPEs, so that an RPE is "near zero" if it is within one standard deviation of its mean. Using the benchmark parameter values listed in the caption for Figure 1, we find that the sample standard deviation of RPEs is 0.14. Accordingly, we set $\delta = 0.14$.

For the values of A_α, B_α, and β_{MB}, we follow Lee, Shimojo, and O'Doherty (2014) and set $\alpha(1) = 0.1$ and $\beta(1) = 0.01$. We also set $B_\alpha = 2$. The condition $\alpha(1) = 0.1$ then implies $A_\alpha = 0.1(1 + \exp(B_\alpha)) = 0.839$. Given that β decreases in χ_{MB}, we know that $\beta(1) = 0.01$ is a lower bound for β_{MB}. We set $\beta_{MB} = 0.2$.

We now simulate data from the above framework and examine its implications. We take the initial weight on the model-based system at time 0 to be 0.75. The remaining parameter values are those listed in the caption for Figure 1; in particular, there are 300,000 investors.

At each date from $t = 1$ to $t = 30$, we compute, for each investor i, his estimate of the reliability of each system and the weight that he assigns to the model-based system, $w^i_{MB,t}$; at each date, we then calculate the mean reliability measure across investors and the mean and dispersion of the weights across investors. The left panel in Figure A2 plots the evolution over time of the mean model-free reliability measure. The solid line in the right panel plots the evolution over time of the mean weight assigned to the model-based system. As an indication of the dispersion across investors in the weight on the model-based system, the dashed lines plot the weights that are one standard deviation away from the mean weight.

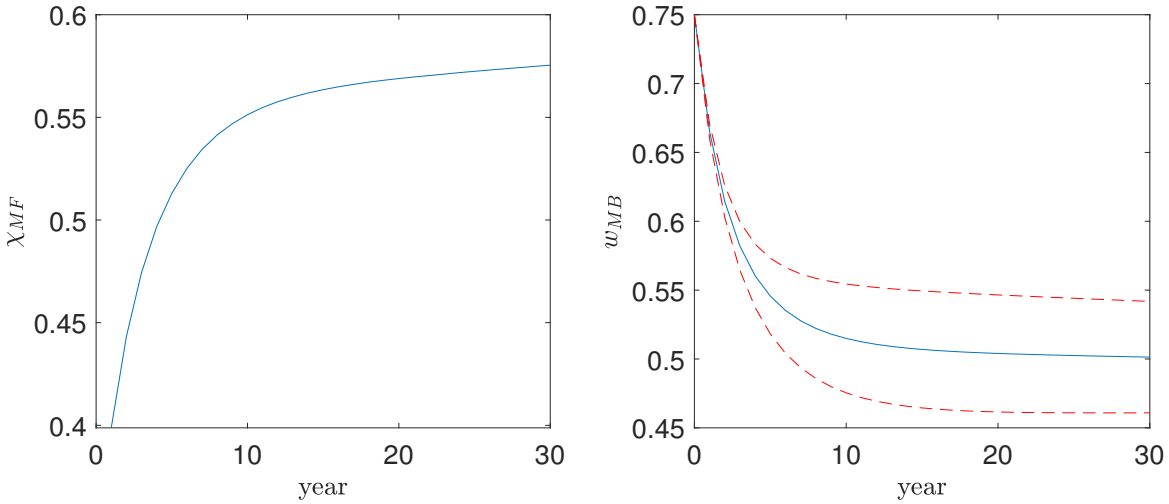

Figure A2. The left graph shows the evolution, over time, of the average investor measure of the reliability of the model-free system. The solid line in the right graph shows the mean weight that investors put on the model-based system. The dashed lines plot the weights on the model-based system that are one standard deviation away from the mean weight. There are 300,000 investors. We set $L = T = 30$, $\alpha^{MF}_{\pm} = \alpha^{MB}_{\pm} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, $w = 0.5$, and $b = 0$, so that there is no generalization.

The left panel shows that, over time, as the model-free system gains experience, it is viewed as more reliable. The right panel shows that, consistent with this, investors over time assign more weight to the model-free system and less to the model-based system.

We conjecture that, if the volatility of recent stock market returns has been high, this will lower the measured reliability of the model-free system, χ_{MF}. To check this, we run the following panel regression with 300,000 investors and 30 years of data:

$$\chi^i_{MF,t} = \alpha + \beta_1 t + \beta_2 \sigma_{t,i} + \varepsilon_{t,i}, \tag{7}$$

where $\sigma_{t,i}$ is the sample standard deviation of stock market returns computed using 10 years of past returns experienced by investor i. We obtain $\beta_1 = 0.004$ and $\beta_2 = -0.581$. These results provide support for our conjecture: the reliability of the model-free system decreases following volatile market returns.

We can also test a related conjecture, which is that the reliability of the model-free system goes down after an extreme market return. We create an indicator variable that equals one if the most recent stock market return deviates from its mean by more than two standard deviations – specifically, if the gross market return exceeds 1.447 or falls below 0.614. We then run the following panel regression with 300,000 investors and 30 years of data:

$$\chi^i_{MF,t} = \alpha + \beta_1 t + \beta_2 \mathbf{1}_{R^i_{m,t} \text{ is extreme}} + \varepsilon_{t,i}. \tag{8}$$

We obtain $\beta_1 = 0.004$ and $\beta_2 = -0.032$. The negative coefficient β_2 confirms that the reliability of the model-free system goes down following an extreme market return.

C. The Relationship between Model-free Allocations and Past Returns: Comparative Statics

The graphs in Figure 3 of the main text show how the relationship between investors' time T model-free allocations and past stock market returns changes as we vary one of the model parameters while keeping the others at their benchmark levels. Across the four graphs, we vary the degree of generalization, the degree of exploration, the discount factor, and the number of allocation choices. Changing these parameters would have little effect on model-*based* allocations. However, Figure 3 shows that it has significant impact on model-free allocations. In this section, we explain the intuition for these patterns.

Generalization. The top-left graph in Figure 3 plots the coefficients in a regression of the time T 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 equation (14) of the main text, normalized as a probability distribution, a standard deviation equal to that of a uniform distribution whose support has a width of 0.2, 0.4, and 0.8, respectively.

The graph shows that, 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 outsized influence

on his later allocations.

Exploration. The top-right graph in Figure 3 plots the relationship between the model-free allocation and past market returns for three different values of β, which controls the degree of exploration, namely 10, 50, and ∞. Recall that, as β rises, the investor explores less: he is more likely to choose the allocation with the highest estimated Q value; when $\beta = \infty$, he always chooses this allocation. We find that, for a wide range of values of β – any β below 80 – the model-free allocation puts positive weights on past returns that decline over most of the time range, as they do for our benchmark case of $\beta = 30$. However, when β is higher than 80, the weights on past returns increase for more than half of the time 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.

Discount factor. The bottom-left graph plots the relationship between the model-free allocation and past market returns for three different values of the discount factor 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 striking 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 factor does not affect the dependence of allocations on past returns.

Number of allocations. In the main text, we allowed investors to select from one of 11 possible allocations. The bottom-right graph in Figure 3 shows how the time T model-free allocation depends on past market returns as we vary the number of allocation options, 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 role. This property of the model-free system again distinguishes it 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 higher 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.

In summary, the model-free system has rich implications for the relationship between allocations and past market returns. While this relationship is typically downward-sloping, it can sometimes be 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 market returns is affected by factors that play little to no role for the model-based system.

D. Analytical Results

While the Q-learning algorithm is simple to state, it is difficult to derive analytical results about its predictions. Nonetheless, for certain 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 factor $\gamma = 0$, 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 market returns. By comparing equations (9) and (18), we confirm that the model-free system puts substantially greater weight on distant past returns.

We then turn to the 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 market returns. Comparing equations (21) and (34)-(35), 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 for the case where the discount factor γ is greater than zero. However, the resulting expressions are messier and do not provide much additional insight.

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 holds:

$$\lim_{t\to\infty} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \frac{\beta R^{2\beta-1}}{(R^\beta + 1)^{k+3}} \tag{9}$$

for $k \geq 0$.

Proof: At any time $t > 0$,

$$\begin{aligned} Q_t(0) &= \log(R_f) = 1, \\ Q_t(1) &= \log(R_{m,t'}), \end{aligned} \tag{10}$$

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

$$\begin{aligned}
\mathbb{E}[a_t] &= \mathbb{P}(a_t = 1) \\
&= \sum_{i=0}^{t-1} \mathbb{P}(a_t = 1 | i \text{ is the largest index s.t. } a_i = 1) \times \mathbb{P}(a_i = 1) \\
&\quad + \mathbb{P}(a_t = 1 | a_0 = \ldots = a_{t-1} = 0) \times \mathbb{P}(a_0 = \ldots = a_{t-1} = 0) \\
&= \left(\sum_{i=0}^{t-1} \frac{R_{m,i+1}^{\beta}}{R_{m,i+1}^{\beta} + 1} \left(\frac{1}{R_{m,i+1}^{\beta} + 1} \right)^{t-i-1} \times \mathbb{P}(a_i = 1) \right) + \frac{1}{2^{t+1}}. \qquad (11)
\end{aligned}$$

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

$$\mathbb{P}(a_t = 1) = \frac{(2^{t+1} - 1)R^{\beta} + 1}{2^{t+1}(R^{\beta} + 1)}, \quad \forall t \geq 0. \qquad (12)$$

The verification of (12) is as follows. When $t = 0$, equation (12) implies that $\mathbb{P}(a_0 = 1) = \frac{1}{2}$, which is clearly true. For $t = j \geq 1$, suppose (12) is true for $0 \leq i \leq j - 1$. Then, from equation (11), we have

$$\begin{aligned}
\mathbb{P}(a_j = 1) &= \left(\sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{j-i}} \times \mathbb{P}(a_i = 1) \right) + \frac{1}{2^{j+1}} \\
&= \left(\sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{j-i}} \times \frac{(2^{i+1} - 1)R^{\beta} + 1}{2^{i+1}(R^{\beta} + 1)} \right) + \frac{1}{2^{j+1}} \\
&= \frac{R^{\beta}(1 - 2^{-j})}{R^{\beta} + 1} + \frac{1}{2^{j+1}} = \frac{(2^{j+1} - 1)R^{\beta} + 1}{2^{j+1}(R^{\beta} + 1)}. \qquad (13)
\end{aligned}$$

That is, (12) is also true for $t = j$.

Equation (12) 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,

$$\begin{aligned}
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t}} &= \frac{\partial \mathbb{P}(a_t = 1)}{\partial R_{m,t}} = \frac{\partial \left[\frac{R_{m,t}^{\beta}}{R_{m,t}^{\beta} + 1} \mathbb{P}(a_{t-1} = 1) \right]}{\partial R_{m,t}} \\
&= \frac{\beta R_{m,t}^{\beta - 1}}{(R_{m,t}^{\beta} + 1)^2} \mathbb{P}(a_{t-1} = 1) = \frac{\beta R^{\beta - 1}}{(R^{\beta} + 1)^2} \frac{(2^t - 1)R^{\beta} + 1}{2^t(R^{\beta} + 1)}. \qquad (14)
\end{aligned}$$

As t goes to infinity, we obtain

$$\lim_{t \to \infty} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t}} = \frac{\beta R^{2\beta - 1}}{(R^{\beta} + 1)^3}, \qquad (15)$$

which is the same as (9) when $k = 0$.

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

$$\begin{aligned}
\frac{\partial \mathbb{P}(a_t = 1)}{\partial R_{m,t-k}} &= \left(\sum_{i=t-k}^{t-1} \frac{R_{m,i+1}^{\beta}}{(R_{m,i+1}^{\beta} + 1)^{t-i}} \cdot \frac{\partial \mathbb{P}(a_i = 1)}{\partial R_{m,t-k}} \right) + \frac{\partial \left[\frac{R_{m,t-k}^{\beta}}{(R_{m,t-k}^{\beta}+1)^{k+1}} \mathbb{P}(a_{t-k-1} = 1) \right]}{\partial R_{m,t-k}} \\
&= \left(\sum_{i=t-k}^{t-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{t-i}} \cdot \frac{\partial \mathbb{P}(a_i = 1)}{\partial R_{m,t-k}} \right) + \frac{\beta R^{\beta-1} - k\beta R^{2\beta-1}}{(R^{\beta} + 1)^{k+2}} \cdot \mathbb{P}(a_{t-k-1} = 1) \\
&= \sum_{i=0}^{k-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{i+1}} \cdot \frac{\partial \mathbb{P}(a_{t-i-1} = 1)}{\partial R_{m,t-k}} \\
&\quad + \frac{\beta R^{\beta-1} - k\beta R^{2\beta-1}}{(R^{\beta} + 1)^{k+2}} \cdot \frac{(2^{t-k} - 1)R^{\beta} + 1}{2^{t-k}(R^{\beta} + 1)}. \qquad (16)
\end{aligned}$$

Suppose (9) is true for $0 \leq k \leq j - 1$. Then

$$\begin{aligned}
\lim_{t\to\infty} \frac{\partial \mathbb{P}(a_t = 1)}{\partial R_{m,t-j}} &= \sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{i+1}} \cdot \lim_{t\to\infty} \frac{\partial \mathbb{P}(a_{t-i-1} = 1)}{\partial R_{m,t-j}} \\
&\quad + \frac{\beta R^{\beta-1} - j\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j+2}} \cdot \frac{R^{\beta}}{R^{\beta} + 1} \\
&= \left(\sum_{i=0}^{j-1} \frac{R^{\beta}}{(R^{\beta} + 1)^{i+1}} \cdot \frac{\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j-i+2}} \right) + \frac{\beta R^{\beta-1} - j\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j+2}} \cdot \frac{R^{\beta}}{R^{\beta} + 1} \\
&= \frac{j\beta R^{3\beta-1}}{(R^{\beta} + 1)^{j+3}} + \frac{\beta R^{2\beta-1} - j\beta R^{3\beta-1}}{(R^{\beta} + 1)^{j+3}} = \frac{\beta R^{2\beta-1}}{(R^{\beta} + 1)^{j+3}}. \qquad (17)
\end{aligned}$$

That is, (9) holds for $k = j$. Equation (17) completes an inductive proof of (9). ■

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 holds:

$$\begin{aligned}
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t}} &= \frac{\beta R_{m,t}^{\beta-1}}{(R_{m,t}^{\beta} + 1)^2}, \\
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} &= 0, \quad k > 0. \qquad (18)
\end{aligned}$$

Proof: At any time $t > 0$,

$$\begin{aligned}
Q_t(0) &= 0, \\
Q_t(1) &= \log(R_{m,t}). \qquad (19)
\end{aligned}$$

The softmax rule implies

$$\mathbb{E}[a_t] = \mathbb{P}(a_t = 1) = \frac{R_{m,t}^{\beta}}{R_{m,t}^{\beta} + 1}. \tag{20}$$

Taking the derivative of (20) with respect to $R_{m,t-k}$ leads to (18). ■

Theorem 3 (Model-free learning): Assume that $\alpha \in (0, 1]$, $\beta > 0$, $\gamma = 0$, $R_f = 1$, and that there are two possible allocations $\{0, 1\}$. Set $Q_0(0) = Q_0(1) = 0$. Assume that $R_{m,i} \equiv R$ for all periods $i \geq 1$. Further assume that, when an investor allocates money to 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 holds:

$$\lim_{t\to\infty} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{2\beta-1}}{(R^{\beta}+1)^3} \left(\frac{R^{\beta} + 1 - \alpha R^{\beta}}{R^{\beta} + 1}\right)^k. \tag{21}$$

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

$$\begin{aligned}
\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} &= \sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t} \frac{\partial \left[\mathbb{P}(a_t = 1|a_i = b_i, \forall i \in [t-1])\mathbb{P}(a_i = b_i, \forall i \in [t-1])\right]}{\partial R_{t-k}} && (22)\\
&= \sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t} \frac{\partial \mathbb{P}(a_t = 1|a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} \mathbb{P}(a_i = b_i, \forall i \in [t-1]) && (23)\\
&\quad + \sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t} \frac{\partial \mathbb{P}(a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} \mathbb{P}(a_t = 1|a_i = b_i, \forall i \in [t-1]). && (24)
\end{aligned}$$

We analyze the expressions in (23) and (24) separately. First, we derive $\lim_{t\to\infty}(23)$, the limit of the expression in (23) as t goes to infinity. We have

$$\frac{\partial \mathbb{P}(a_t = 1|a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} = \frac{\partial \left(\frac{e^{\beta Q_t(1)}}{e^{\beta Q_t(1)}+1}\right)}{\partial R_{t-k}} = \frac{1}{(e^{\beta Q_t(1)} + 1)^2} \frac{\partial e^{\beta Q_t(1)}}{\partial R_{t-k}}. \tag{25}$$

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 note that $\frac{1}{(e^{\beta Q_t(1)}+1)^2} = \frac{1}{(R^{\beta}+1)^2}$, 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 (25), 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,

$$\frac{\partial e^{\beta Q_t(1)}}{\partial R_{t-k}} = \frac{(1-\alpha)^n \beta e^{\beta Q_t(1)}}{R} = (1-\alpha)^n \beta R^{\beta-1} \tag{26}$$

and (25) can be simplified as

$$\frac{\partial \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} = \frac{(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta + 1)^2}. \tag{27}$$

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, (25) can be simplified as

$$\frac{\partial \mathbb{P}(a_t = 1 | a_i = b_i, \forall i \in [t-1])}{\partial R_{t-k}} = \frac{\alpha(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta + 1)^2}. \tag{28}$$

Substituting (27) and (28) back into (23), we obtain

$$\begin{aligned} (23) \;=\; & \sum_{n=0}^{k} \sum_{\substack{(b_{t-k},\ldots,b_{t-1}) \in (0,1)^k \\ \sum_{j=t-k}^{j=t-1} b_j = n,\, b_{t-k-1}=1}} \frac{(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta+1)^2} \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) = (0,\ldots,0)}\right) \\ & + \sum_{n=0}^{k} \sum_{\substack{(b_{t-k},\ldots,b_{t-1}) \in (0,1)^k \\ \sum_{j=t-k}^{j=t-1} b_j = n,\, b_{t-k-1}=1}} \frac{\alpha(1-\alpha)^n \beta R^{\beta-1}}{(R^\beta+1)^2} \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) \neq (0,\ldots,0)}\right). \end{aligned} \tag{29}$$

Note that

$$0 \leq \mathbb{P}\left(\substack{a_i = b_i, \forall i \in [t-k-1, t-1], \\ (a_0,\ldots,a_{t-k-2}) = (0,\ldots,0)}\right) \leq \mathbb{P}((a_0, \ldots, a_{t-k-2}) = (0, \ldots, 0)) = \frac{1}{2^{t-k-1}}. \tag{30}$$

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

$$\begin{aligned} \mathbb{P}(a_t = 1) \;=\; & \mathbb{P}(a_t = 1 | (a_0, \ldots, a_{t-1}) = (0, \ldots, 0)) \cdot \mathbb{P}((a_0, \ldots, a_{t-1}) = (0, \ldots, 0)) \\ & + \mathbb{P}(a_t = 1 | (a_0, \ldots, a_{t-1}) \neq (0, \ldots, 0)) \cdot \mathbb{P}((a_0, \ldots, a_{t-1}) \neq (0, \ldots, 0)) \\ \;=\; & \frac{1}{2}\left(\frac{1}{2}\right)^t + \frac{R^\beta}{R^\beta + 1}\left(1 - \left(\frac{1}{2}\right)^t\right), \end{aligned} \tag{31}$$

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

$$
\begin{aligned}
&\lim_{t\to\infty}(23)\\
=\;&\sum_{n=0}^{k}\frac{\alpha(1-\alpha)^n\beta R^{\beta-1}}{(R^\beta+1)^2}\lim_{t\to\infty}\sum_{\substack{(b_{t-k},\ldots,b_{t-1})\in(0,1)^k\\ \sum_{j=t-k}^{j=t-1}b_j=n,\,b_{t-k-1}=1}}\mathbb{P}\big(\substack{a_i=b_i,\forall i\in[t-k-1,t-1],\\(a_0,\ldots,a_{t-k-2})\neq(0,\ldots,0)}\big)\\
=\;&\sum_{n=0}^{k}\frac{\alpha(1-\alpha)^n\beta R^{\beta-1}}{(R^\beta+1)^2}\Big(\lim_{t\to\infty}\mathbb{P}(a_{t-k-1}=1)\Big)\lim_{t\to\infty}\sum_{\substack{(b_{t-k},\ldots,b_{t-1})\in(0,1)^k\\ \sum_{j=t-k}^{j=t-1}b_j=n}}\mathbb{P}\big(a_i=b_i,\forall i\in[t-k,t-1]|a_{t-k-1}=1\big)\\
=\;&\sum_{n=0}^{k}\frac{\alpha(1-\alpha)^n\beta R^{\beta-1}}{(R^\beta+1)^2}\frac{R^\beta}{R^\beta+1}\binom{k}{n}\left(\frac{R^\beta}{R^\beta+1}\right)^n\left(\frac{1}{R^\beta+1}\right)^{k-n}\\
=\;&\frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^{3+k}}\sum_{n=0}^{k}\binom{k}{n}(1-\alpha)^nR^{n\beta}\\
=\;&\frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^{3+k}}(1+(1-\alpha)R^\beta)^k=\frac{\alpha\beta R^{2\beta-1}}{(R^\beta+1)^3}\left(\frac{R^\beta+1-\alpha R^\beta}{R^\beta+1}\right)^k. \qquad (32)
\end{aligned}
$$

We now turn to (24). We have

$$
\begin{aligned}
(24)\;=\;&\sum_{\substack{(b_0,\ldots,b_{t-1})\in\{0,1\}^t\\(b_0,\ldots,b_{t-1})\neq(0,\ldots,0)}}\frac{\partial\mathbb{P}(a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}}\frac{R^\beta}{R^\beta+1}\\
&+\frac{\mathbb{P}((a_0,\ldots,a_{t-1})=(0,\ldots,0))}{\partial R_{t-k}}\cdot\frac{1}{2}\\
=\;&\sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t}\frac{\partial\mathbb{P}(a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}}\frac{R^\beta}{R^\beta+1}\\
&+\frac{\mathbb{P}((a_0,\ldots,a_{t-1})=(0,\ldots,0))}{\partial R_{t-k}}\left(\frac{1}{2}-\frac{R^\beta}{R^\beta+1}\right)\\
=\;&\frac{\partial\sum_{(b_0,\ldots,b_{t-1})\in\{0,1\}^t}\mathbb{P}(a_i=b_i,\forall i\in[t-1])}{\partial R_{t-k}}\frac{R^\beta}{R^\beta+1}\\
=\;&0. \qquad (33)
\end{aligned}
$$

Finally, (32) and (33) together lead to (21). ■

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

Given these assumptions,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} = \frac{\alpha\beta R^{\beta-1}}{(R^\beta+1)^2}(1-\alpha)^k \tag{34}$$

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

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,1}} = \frac{\beta R^{\beta-1}}{(R^\beta+1)^2}(1-\alpha)^{t-1}. \tag{35}$$

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

$$\begin{aligned} Q_t(1) - Q_t(0) &= \mathbb{E}_t^p(R_{m,t+1}) \\ &= (1-\alpha)^{t-1}\log(R_{m,1}) + \alpha\sum_{j=2}^{t}(1-\alpha)^{t-j}R_{m,j} \\ &= \log(R). \end{aligned} \tag{36}$$

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

$$\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,t-k}} = \frac{\alpha(1-\alpha)^k}{R}, \tag{37}$$

and for $k = t-1$,

$$\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,1}} = \frac{(1-\alpha)^{t-1}}{R}. \tag{38}$$

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

$$\begin{aligned} \frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} &= \frac{\partial\left(\frac{e^{\beta(Q_t(1)-Q_t(0))}}{e^{\beta(Q_t(1)-Q_t(0))}+1}\right)}{\partial R_{m,t-k}} \\ &= \frac{\beta e^{\beta(Q_t(1)-Q_t(0))}}{(e^{\beta(Q_t(1)-Q_t(0))}+1)^2}\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,t-k}} \\ &= \frac{\beta R^\beta}{(R^\beta+1)^2}\frac{\partial\left(Q_t(1)-Q_t(0)\right)}{\partial R_{m,t-k}}. \end{aligned} \tag{39}$$

Substituting (37) and (38) into (39) then gives (34) and (35), respectively. ■

E. Parameter Estimation

In this section, we describe in more detail the procedure that we use in Section 4.7 to estimate the values of four important parameters in our framework: 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 the weight w on the model-based system. We do the estimation in two steps. We first use data on investor beliefs to estimate $\bar{\alpha}^{MB}$. We then estimate $\bar{\alpha}^{MF}$, β, and w by targeting two facts discussed in Section 4 of the paper, namely the sensitivity of allocations to beliefs in Giglio et al. (2021) and the experience effect in Malmendier

and Nagel (2011). We keep the remaining parameters at their benchmark values, 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 market returns. Specifically, as in Greenwood and Shleifer (2014), we take monthly Gallup data from October 1996 to November 2011 on average investor beliefs about future one-year stock market returns and regress these beliefs on past annual stock market returns. The coefficients on the returns one, two, and three years in the past are 0.127, 0.037, and 0.029, respectively; the ratio of the second coefficient to the first is 0.29 and the ratio of the third coefficient to the second is 0.77. We search for a value of $\bar{\alpha}_{MB}$ that, in simulated data from the model-based system, best matches the first coefficient, 0.127, and the two subsequent rates of decline in the coefficients, 0.29 and 0.77; intuitively, we are trying to match the level and slope of the relationship between beliefs and returns.

To do this, we take $300,000$ investors in six cohorts of $50,000$ 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, α_+^{MB} and α_-^{MB}, from a uniform distribution centered at $\bar{\alpha}_{MB}$ and with width $\Delta = 0.5$. We then compute investor beliefs at each time, as determined by the model-based system and in particular by equations (17) and (18) in the main text. 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 coefficients c_1, c_2, and c_3 on the annual returns one, two, and three years in the past, respectively. We repeat this exercise for many different values of $\bar{\alpha}^{MB}$ and select the value of $\bar{\alpha}^{MB}$ that minimizes

$$(c_1 - 0.127)^2 + (\frac{c_2}{c_1} - 0.29)^2 + (\frac{c_3}{c_2} - 0.77)^2. \tag{40}$$

We find this to be $\bar{\alpha}^{MB} = 0.33$.

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. For given values of $\bar{\alpha}^{MF}$, β, and w, we can compute this coefficient, d, in simulated data from our framework. Due to computational constraints, these data are now based on 60,000 investors in six cohorts of 10,000 each.

Our second target is the functional form in (28) in the main text with $\lambda = 1.3$, which Malmendier and Nagel (2011) use to capture empirical experience effects. Intuitively, we are looking for parameter values that minimize the distance between unnormalized versions of the solid and the dashed lines in the six graphs in Figure 6. 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 another vector of 30 coefficients given by

$$0.972\frac{(31-j)^{1.3}}{\sum_{l=1}^{30}(31-l)^{1.3}}, \qquad j = 1, 2, \ldots, 30,$$

which, according to column (i) in Table IV of Malmendier and Nagel (2011), captures the empirical relationship between allocations and returns j years in the past for a cohort of age 30. We then compute the L^2 norm of the difference between the two vectors and 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 values of $\{\bar{\alpha}^{MF}, \beta, w\}$. In particular, for many values of $\{\bar{\alpha}^{MF}, \beta, w\}$, we compute

$$(d-1)^2 + \sum_{i=1}^{6} \text{MSE}_i \tag{41}$$

and identify the parameter values that minimize this quantity. The first term in (41) targets the empirical sensitivity of allocations to beliefs, while the second term targets the empirical experience effect. We find that the parameter values that minimize (41) are $\bar{\alpha}^{MF} = 0.26$, $\beta = 20$, and $w = 0.38$.

F. Allowing for State Dependence

In the main text, we focus on the case with no state dependence and find that this case already delivers a rich set of results. In this section, we show how an explicit state dependence can be incorporated into our framework. In particular, we consider a setting introduced in Section 5 of the paper in which, rather than having an i.i.d return distribution, the risky asset has state-dependent returns that capture long-run mean-reversion.

Specifically, at each point in time t, we define the recent trend of asset returns as

$$S_{m,t} = (1-\theta) \sum_{i=0}^{t-1} \theta^i R_{m,t-i} + \theta^t S_{m,0}, \tag{42}$$

where $0 < \theta < 1$ is a decay parameter and $S_{m,0}$ is the initial level of the trend at $t = 0$. We specify asset returns so that a good past trend is followed, on average, by low returns, and a bad trend is followed, on average, by high returns. Formally, if $S_{m,t} > \overline{S}$, the next period's return is governed by

$$\log R_{m,t+1} = \mu_L + \sigma \varepsilon_{t+1}, \tag{43}$$

where μ_L has a low value; we call this the Low state, L. If, on the other hand, $S_{m,t} < \underline{S}$, the next period's return is governed by

$$\log R_{m,t+1} = \mu_H + \sigma \varepsilon_{t+1}, \tag{44}$$

where μ_H has a high value; we call this the High state, H. Finally, if $\underline{S} \leq S_{m,t} \leq \overline{S}$, the next period's return is governed by

$$\log R_{m,t+1} = \mu_M + \sigma \varepsilon_{t+1}, \tag{45}$$

where μ_M takes a moderate value; we call this the Moderate state, M. In each case, ε_{t+1} is drawn from a standard Normal distribution, independently of other shocks.

If an investor fails to recognize the existence of the three market states, L, M, and H,

then, to update his Q values, $Q_t^{MF}(a)$ and $Q_t^{MB}(a)$, he follows the model-free and model-based algorithms described in Sections 2.2 and 2.3 of the main text. If the investor *is* instead able to recognize and observe the three states, his learning algorithms are different. For model-free learning, the Q values are updated according to

$$Q_{t+1}^{MF}(s_t, a) = Q_t^{MF}(s_t, a) + \alpha_{t,\pm}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(s_{t+1}, a') - Q_t^{MF}(s_t, a)] \tag{46}$$

at time $t+1$, where s_t and s_{t+1} can be L, M, or H. For simplicity, we do not consider generalization.

For model-based learning, following a market return $R_{m,t+1} = R$, the probability estimates are updated according to

$$p_{t+1}(R_m = R, s_t) = p_t(R_m = R, s_t) + \alpha_{t,\pm}^{MB}[1 - p_t(R_m = R, s_t)] \tag{47}$$

at time $t+1$; the learning rate $\alpha_{t,+}^{MB}$ applies when $R > 1$ and the learning rate $\alpha_{t,-}^{MB}$ applies when $R \leq 1$. These probability estimates allow the investor to perceive three return distributions, one for each state. We define the model-based Q values at time t as follows:

$$\begin{aligned}
Q_t^{MB}(s_t = L, a) &= \mathbb{E}_t^{p,L} \log((1-a)R_f + aR_{m,t+1}) + \gamma(\chi^{LH}V^H + \chi^{LM}V^M + \chi^{LL}V^L), \\
Q_t^{MB}(s_t = M, a) &= \mathbb{E}_t^{p,M} \log((1-a)R_f + aR_{m,t+1}) + \gamma(\chi^{MH}V^H + \chi^{MM}V^M + \chi^{ML}V^L), \\
Q_t^{MB}(s_t = H, a) &= \mathbb{E}_t^{p,H} \log((1-a)R_f + aR_{m,t+1}) + \gamma(\chi^{HH}V^H + \chi^{HM}V^M + \chi^{HL}V^L),
\end{aligned} \tag{48}$$

where $\mathbb{E}_t^{p,s}$ represents the investor's perceived return distribution in state s at time t, χ^{s_1,s_2} represents the investor's perceived transition probability from state s_1 at time t to state s_2 at time $t+1$, and V^s represents the investor's perceived optimal valuation of state s.

The hybrid Q values are

$$Q_t^{HYB}(s_t, a) = (1-w)Q_t^{MF}(s_t, a) + wQ_t^{MB}(s_t, a). \tag{49}$$

Finally, the investor chooses her portfolio allocation probabilistically, according to

$$p(s_t, a_t = a) = \frac{\exp[\beta Q_t^{HYB}(s_t, a)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(s_t, a')]}. \tag{50}$$

Equation (50) shows that the values of χ^{s_1,s_2} and V^s do not affect the investor's allocation choice: within each state, the part of Q_t^{HYB} in the numerator of equation (50) that contains χ^{s_1,s_2} and V^s is cancelled out by the same term in the denominator.

We now present some numerical analysis. The parameters σ, $\alpha_{t,\pm}^{MF}$, $\alpha_{t,\pm}^{MB}$, γ, w, and β take the baseline values used in Figure 1 of the paper. In addition, we set $\theta = 0.8$, $\mu_H = 6\%$, $\mu_M = 1\%$, $\mu_L = -4\%$, $\overline{S} = \exp(\mu_M + 0.5\sigma^2) + 3\% = 1.0605$, and $\underline{S} = \exp(\mu_M + 0.5\sigma^2) - 3\% = 1.0005$. The simulation setup is the same as in Figure 1; in particular, there are 300,000 investors. We consider two cases: the case where investors do not recognize the existence of the three states, and the case where they do recognize and observe the three states. In each case, we study the performance and recommended allocations of the model-free system, the model-based system, and the hybrid system. To evaluate performance, we look at each

investor's excess portfolio return from t to $t+1$, where t goes from 0 to 29; we compute the mean and standard deviation of these 30 excess returns for each investor; finally, we average these numbers across the 300,000 investors. To study allocations, we look at each investor's portfolio allocation at time 30; we then average these allocations across the investors who are facing an asset that is in state s at time 30, where s is L, M, or H.

The table below presents the performance measures and allocations for the model-free, model-based, and hybrid systems in the case where the algorithms do not recognize the existence of the three states:

	mean	stdev	$\overline{a}_L$	$\overline{a}_M$	$\overline{a}_H$
MF	1.61%	12.99%	60.53%	57.50%	49.79%
MB	0.96%	12.77%	75.24%	52.54%	29.06%
hybrid	1.20%	12.67%	67.59%	53.73%	37.18%

The table below presents the performance measures and allocations for the model-free, model-based, and hybrid systems in the case where the algorithms do recognize the existence of the three states:

	mean	stdev	$\overline{a}_L$	$\overline{a}_M$	$\overline{a}_H$
MF	1.77%	12.98%	49.15%	53.03%	59.41%
MB	1.94%	13.25%	41.97%	51.99%	61.69%
hybrid	1.89%	13.05%	43.04%	52.23%	62.77%

We make three observations about these results.

As noted in Section 5.1 of the paper, when investors do not recognize the three market states, the model-free system significantly outperforms the model-based system: the mean excess portfolio return is 1.61% for the model-free system but only 0.96% for the model-based system, while the standard deviation of portfolio returns is similar for the two systems. As shown in Section 3 of the paper, the model-free system is less extrapolative than the model-based system, and this is valuable when there is mean-reversion in asset returns.

When investors do recognize the three market states, the two systems have fairly similar performance: the mean excess portfolio return is 1.94% for the model-based system and 1.77% for the model-free system. On the one hand, the slow learning of the model-free system means that this system is slower to recognize the lower (higher) returns in the Low (High) state, which is costly. At the same time, this system also exhibits a less extrapolative asset demand, which is beneficial.

When the model-based system is able to recognize the three market states but the model-free system is not, a tension arises between the two systems, as suggested in Section 5.2. Following a sequence of good returns, the model-free system recommends a high allocation: when $S_{m,t} > \overline{S}$, the average allocation recommended by the state-independent model-free system is 60.53%, higher than what it recommends in the Moderate state. By contrast,

the model-based system recognizes that a good trend is often followed by low returns and hence recommends a low allocation: when $S_{m,t} > \overline{S}$, the average allocation recommended by the state-dependent model-based system is 41.97%, lower than what it recommends in the Moderate state. One system therefore pulls the investor toward a higher allocation in the Low state, while the other pulls him toward a lower allocation.

G. SARSA: An Alternative Model-free Framework

The model-free frameworks most widely used by psychologists are Q-learning and SARSA. In the main text, 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. At time 0, all Q values are set to zero: $Q_0^{MF}(a) = 0$, $\forall a$. The investor chooses one of the possible 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

$$p(a_t = a) = \frac{\exp[\beta Q_{t-1}^{MF}(a)]}{\sum_{a'} \exp[\beta Q_{t-1}^{MF}(a')]}, \tag{51}$$

and given $R_{p,t}$ and a_t, he updates the Q value of his previous allocation a_{t-1} from $Q_{t-1}^{MF}(a_{t-1})$ to $Q_t^{MF}(a_{t-1})$ according to

$$Q_t^{MF}(a_{t-1}) = Q_{t-1}^{MF}(a_{t-1}) + \alpha_{t-1,\pm}^{MF} \left[\log R_{p,t} + \gamma Q_{t-1}^{MF}(a_t) - Q_{t-1}^{MF}(a_{t-1})\right]. \tag{52}$$

Analogous to the analysis in Section 3.2, we examine how investors' date T allocations a_T recommended by each of SARSA, Q-learning, and model-based learning depend on the past market returns investors have been exposed to. Figure A3 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, while the allocation recommended by model-based learning depends primarily on recent past returns, the allocations recommended by both Q-learning and SARSA depend significantly even on distant past returns.

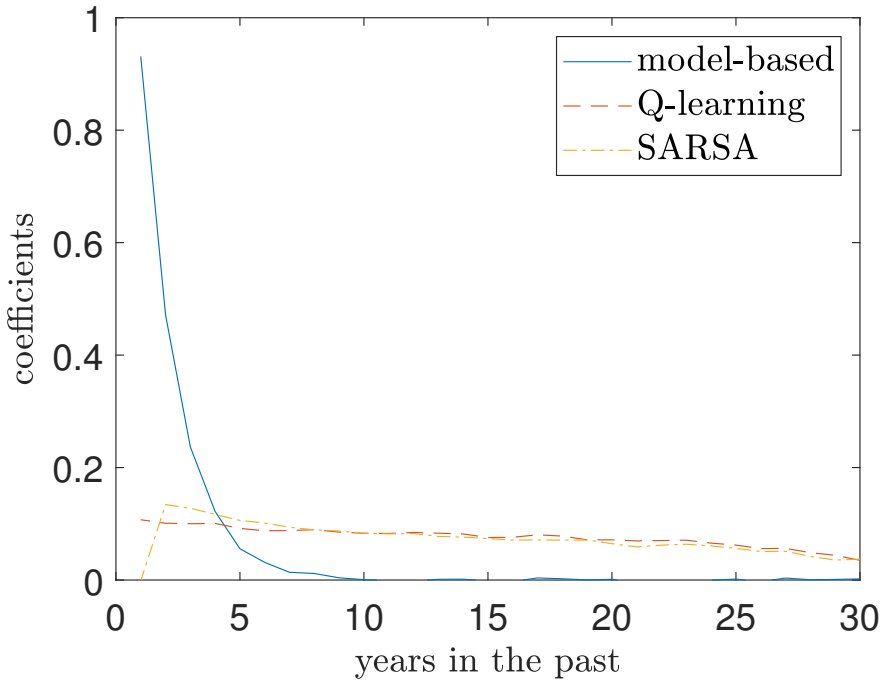

Figure A3. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients for three cases: model-based learning; model-free Q-learning; and model-free SARSA. 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.