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

Nicholas Barberis and Lawrence Jin

February 2026*

**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, the insensitivity of allocations to beliefs, the inertia in household allocations, and the dispersion in these allocations. Our analysis suggests that model-free learning may play 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, for very useful feedback, 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 to seminar participants at several universities, including Caltech, Columbia University, Cornell University, Harvard University, MIT, the University of Chicago, the University of Pennsylvania, and Yale University. We are also grateful to Colin Camerer, Nathaniel Daw, Peter Dayan, Samuel Gershman, John O'Doherty, and members of their lab groups for very helpful discussions about the cognitive science underlying the paper. Steven Ma and Youxuan Xu provided outstanding 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, study its implications for investor behavior, and show that it is helpful for thinking about a range of empirical facts.[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 over future outcomes. Rather, it learns from 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 over future outcomes – and then uses this to compute the value of different actions. We use a model-based approach that is often adopted in research in psychology and that, like the model-free system, has

[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). All of these papers are authored by cognitive scientists – by psychologists and neuroscientists. We discuss the behavioral and neural evidence for the framework in more detail in Section 2.

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; 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, which we think of as the stock market, 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 then 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; this is because the beliefs about future returns generated by the model-based system themselves put positive and declining weights on past returns. We find that the model-free system also recommends an allocation that puts positive weights on past returns, and show that it does so through a mechanism that is new to financial economics and has nothing to do with beliefs. In brief: If an investor has a high allocation to the stock market and the market then posts a high return, the high allocation will be strongly rewarded, making it likely that the investor will persist with a high allocation. Conversely, if an investor has a high allocation to the stock market and the market then drops sharply, the high allocation will be negatively rewarded, making it likely that the investor will switch to a lower allocation.

We also find that, while the allocations generated by both the model-free and model-based systems put positive and declining weights on past returns, the decline is much slower in the case of the model-free system, with the result that this system puts much more weight on distant past returns than does the model-based system. This is because the model-free system 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 continue to depend only on more distant past returns. Consequently, it takes a long time for the influence of distant past returns to fade.

Our results imply that the model-free system can offer a new foundation for the notion of "extrapolative demand," the idea, motivated by empirical evidence, that investors' demand for a risky asset depends on a weighted average of the asset's past returns, where the weights are positive and larger for recent returns. While prior work has typically assumed that this extrapolative demand is driven by extrapolative beliefs, our analysis shows that it also emerges from a foundation that has nothing to do with beliefs, namely model-free learning.

We present five applications of our framework, all related to investor behavior and investor beliefs. These are: experience effects; the disconnect between allocations and beliefs in the frequency domain; the insensitivity of allocations to beliefs; inertia in household allocations; and dispersion in these allocations. This list of field applications is striking because, in prior research, our framework has been used almost exclusively to explain behavior in experimental settings. We link our framework to these applications primarily through numerical analysis: we consider a large number of different parameterizations of our framework and show that the empirical phenomenon in question emerges for all, or almost all, of the parameterizations that put significant weight on model-free learning. We complement this numerical analysis with new theoretical results that we obtain in a simplified version of the framework.

We briefly summarize the five applications here. Our framework 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 basic 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.

Our framework can also address a puzzling disconnect between investor beliefs and stock market allocations. While individual investors' beliefs about future stock market returns depend primarily on market returns in the recent past, Malmendier and Nagel (2011) find that these investors' allocations to the stock market depend significantly even on market returns in the distant past. 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 over 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 gives an answer that depends primarily on recent past returns: we noted above that the model-based system puts heavy weight on recent returns. However, the individual's allocation is influenced by both the model-based *and* model-free systems; since the model-free system puts substantial weight even on distant past returns, his allocation does too.

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 less weight on recent returns.

Inertia is a general consequence of model-free learning: as an individual tries different actions over time, one of them may be rewarded with a good outcome, leading the individual to stick with that action going forward. This idea can be applied in many domains. In our

[2] See Ameriks et al. (2020), Giglio et al. (2021), Charles, Frydman, and Kilic (2024), and Yang (2025).

setting, it sheds light on the inertia in household allocations to the stock market. After a household tries an allocation, the allocation may be rewarded with a good portfolio return, leading the household to stick with it in the future. We also show that the model-free system generates substantial dispersion across households in their allocations to the stock market. This offers a new way of understanding the empirically-observed dispersion, one that is not based on differences in objective functions or beliefs, but instead on prior allocations being rewarded with good returns.

Beyond our investigation of the above five applications, we present several additional analyses: We study the predictions of more fully rational versions of the model-free and model-based systems. We consider alternative action spaces beyond percentage allocations to the stock market. We compare the framework's implications to those of more traditional models of inattention. We gauge the framework's flexibility by computing measures of its "completeness" and "restrictiveness." And finally, we lay out its key predictions. The immediate prediction is that model-free estimates of action values will explain allocations, but this is not implementable because the action values are not directly observed. We therefore develop proxies for these action values that are, in principle, observable.

The full name of model-free learning is model-free reinforcement learning. Reinforcement learning is a fundamental concept in both psychology and neuroscience, but has a much smaller footprint in economics and finance. This is a natural moment to revisit reinforcement learning in economics, for at least three reasons: cognitive scientists have developed computational models of human reinforcement learning that we can apply; there is mounting neural evidence for these models; and most important for our purposes, there is now a framework – the framework we use in this paper – that combines reinforcement learning with traditional model-based approaches. Taken alone, reinforcement learning is too extreme for most economic settings; but in combination with model-based learning, it becomes more useful, as we show in this paper.[3]

Our paper is also part of a new wave of research in behavioral economics that moves beyond the work on judgment and decision-making made famous by Daniel Kahneman and

[3] 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.

Amos Tversky, and instead incorporates deeper cognitive foundations into economics, whether about memory (Bordalo, Gennaioli, and Shleifer, 2020), attention (Gabaix, 2019), cognitive noise (Khaw, Li, and Woodford, 2021; Frydman and Jin, 2022; Enke and Graeber, 2023), or, as in this paper, learning algorithms. A hallmark of these cognitive processes is that, in the face of a complex world, they often simplify or approximate. So it is with model-free learning, which is a simple algorithm for making decisions in a complex world. In the long run, it can compute correct action values, but in the shorter term, it leads to departures from full rationality, either because convergence to correct values takes a long time, or because it is not perfectly suited to the environment at hand.

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 study their implications for investor behavior, focusing on how the stock market allocations they recommend depend on past stock market returns. In Section 4, we discuss five applications of our framework to understanding investor allocations and beliefs. Section 5 summarizes some additional analyses 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 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 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:

$$\max_{\{a_t\}_{t=0}^{\infty}} E_0 \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \right], \tag{1}$$

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. A prominent 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 $\alpha_t^{MF}$ 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. Importantly, only the $Q$ value of the most recently-chosen action, $a$, is updated; the $Q$ values of the other actions remain unchanged.

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

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:

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

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 $\beta$ 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 eventually 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

[5] The softmax expression in (4) can also be interpreted in terms of Luce-style random utility. In a random utility specification, even if the values of different options are known, there may be unobserved drivers of choice or plain errors that can be modeled as stochastic choice. The concept of exploration instead recognizes that the action values are estimated imprecisely and that exploration is needed to make them more accurate.

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 are an even bigger factor behind the surge of interest in model-free learning. 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 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

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

where $\alpha^{MF}_{t,\pm} = \alpha^{MF}_{t,+}$ if the reward prediction error is positive and $\alpha^{MF}_{t,\pm} = \alpha^{MF}_{t,-}$ otherwise. For the sake of psychological realism, we also adopt this modification, although it is not required for the applications we discuss later.

## 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

[6] Other experimental studies with a similar theme are Charness and Levin (2005), which we come back to later in Section 2, Payzan-LeNestour and Bossaerts (2015), and Allos-Ferrer and Garagnani (2023).

[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

$$\max_{\{a_t\}_{t=0}^{\infty}} E_0 \left[ \sum_{t=1}^{\infty} \gamma^{t-1} \log R_{p,t} \right], \tag{7}$$

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), there are at least two reasons to think that a model-free system like Q-learning will play a role in his decision-making. First, the model-free system is a fundamental component of human decision-making. As such, it is likely to play a role in any decision unless explicitly "switched off" – and because it operates below the level of conscious awareness, many investors will not recognize its influence and will therefore fail to turn it off. Second, many people do not have a good "model" of financial markets – for example, they have a poor sense of the structure of asset returns. As a consequence, the brain is likely to assign at least some control of financial decision-making to the model-free system – again, without a person's conscious awareness – precisely because this system does not need a model of the environment. This motivates 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

[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.

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

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

– 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

[10] As noted in Section 2.1, another possible interpretation of the probabilistic choice in (11) is Luce-style random utility.

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 $Q$ 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). While such generalization is not required for any of the applications we discuss later, for the sake of psychological realism, we incorporate it 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 of 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

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

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 $\alpha^{MF}_{t,\pm}\kappa(\widehat{a})$, 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]

[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

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 $a + 10\%$. 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]

While we are applying model-free learning in one particular setting, it can be used in a wide range of environments. It can handle any objective function of the form in (1); a complex set of actions, including, for example, allocations to multiple assets; and a rich state structure.

One question that arises in the case of multiple assets is whether the investor's allocation under model-free learning depends on the specific action space. Consider a financial market with two risky assets, $A$ and $B$, and denote a 50:50 mix of the two assets as the "market portfolio." If $a_A$, $a_B$, and $a_M$ are the investor's allocations to the two risky assets and to the market, respectively, we can ask if the investor's allocations to the two assets under model-free learning depend on whether the action space is $\{a_A, a_B\}$, $\{a_A, a_M\}$, or $\{a_B, a_M\}$. In Internet Appendix B, we show that the answer is no: at each point in time, the investor's allocations to assets $A$ and $B$ are identical across the three action spaces.

The decision problem we study here brings to mind the analysis of multi-armed bandits in the field of operations research. In bandit problems, an individual must choose among various options to maximize expected reward; for each option, he does not know the distribution of outcomes and can learn it only by trying the option and observing the outcome. Research on these problems has focused on developing algorithms to guide decision-making, and on proving results about the efficacy of these algorithms (Lattimore and Szepesvari, 2020). In Section 5.1, where we examine more fully rational versions of model-free learning, we will draw

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.

[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.

inspiration from these algorithms. In Sections 2 to 4, however, where our focus is explaining observed behavior, we root our analysis in psychology research and specifically in algorithms that, based on neural evidence, the brain actually appears to use.

## 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 over future outcomes – for example, in our setting, a probability distribution over stock market returns. There are various possible model-based systems. Which one should we adopt? 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 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 $\alpha_t^{MB}$ 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 one. Intuitively, after an outcome occurs, the model-based system increases the probability it assigns to that outcome. 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:

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

In this case of a constant learning rate, the model-based system generates a form of extrapolative beliefs: the investor's expected stock market return at any moment puts weights on past returns that are positive and that decline for more distant past returns.

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, $\alpha_+^{MF}$ and $\alpha_-^{MF}$, for positive and negative reward prediction errors, respectively. Although it is not necessary for the applications we discuss later, for the sake of psychological realism, we extend the model-based algorithm in a similar way, allowing for different model-based learning rates, $\alpha_+^{MB}$ and $\alpha_-^{MB}$, for positive and negative

[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.

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 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.

Charness and Levin (2005) provide experimental evidence for model-free learning in a setting that exhibits a similar tension to the one we just described. In this setting, there are two possible actions, "Left" and "Right." If, after action "Left," the participant receives a reward, this coincides with receiving a signal that, in the next period, it is better to switch to action "Right." And yet, in nearly half of all decisions, participants stick with "Left," the action that was rewarded in the first period – behavior that is very consistent with model-free learning.[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 and, to date, have the largest body of neural evidence in their favor; 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 accounts for several important 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, another experimental setting that generates this tension between model-free and model-based learning is the two-step task; there, too, there is strong evidence for model-free 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. Research in psychology is actively exploring the idea that $w$ varies over time. One hypothesis is that, at each moment, the brain puts more weight on the system it deems more "reliable" at that point (Daw, Niv, and Dayan, 2005). While there is evidence to support this idea, there is as yet no consensus on it, so we do not pursue it further for now.[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, $\{R_{m,-L+1}, \dots, R_{m,0}\}$; uses these to form a perceived distribution of market returns as in (17) and (18); and then computes model-based $Q$ values by way of that distribution, as in (19).[16]

[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.

[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

In Internet Appendix C, we present an example to illustrate the mechanics of the model-free and model-based systems. Specifically, Table A1 shows the $Q$ values that each of an investor's model-free and model-based systems assigns to the 11 possible stock market allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$ over the investor's first six dates of participation in financial markets. Even from a quick glance at the table, we see a key difference between the two systems: at each time, the model-free system primarily updates only the $Q$ value of the most recently-chosen action, while the model-based system updates all 11 $Q$ values based on its currently-perceived stock market return distribution.

# 3 Properties of Investor Behavior

In this section, we study the properties of investor behavior when investors make decisions according to the framework of Section 2. Our focus is on how the allocations recommended by the model-free and model-based systems depend on past stock market returns. 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]

At each date, we allow the investor to choose his stock market allocation $a_t$ from one of 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$. Later in this section, we also consider finer and coarser versions of this allocation set; and in Section 5.5, we consider alternative action spaces – for example, one where the investor chooses the number of shares of the stock market

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.

that he wants to hold.

The four learning rates – $\alpha_+^{MF}$, $\alpha_-^{MF}$, $\alpha_+^{MB}$, and $\alpha_-^{MB}$ – 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 $t$ 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

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

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 learning rates. Later, we allow for dispersion in these rates across investors.[19]

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.

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

[19] One reason 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.

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.[20]

We start by illustrating our results for one particular parameterization of our framework. We will then show that the observed pattern is very robust, in that it also emerges for a wide range of other parameterizations. For the specific parameterization we start with, the parameter values are as follows. As above, $L = T = 30$. Investors' learning rates are set to $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$. The exploration parameter $\beta$ is 30; in simulations, we find that, for this value of $\beta$, 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. We set the discount factor $\gamma$ to 0.97 – this corresponds to an expected investment horizon of 33 years – and we simulate stock market returns from the distribution in (6) with $\mu = 0.01$ and $\sigma = 0.2$; these values provide an approximate fit to historical annual stock market returns. For ease of interpretation, we turn off generalization for now, so that $b = 0$.[21] Finally, we set $w = 0.5$, so that the hybrid system puts equal weight on the model-free and model-based systems.[22]

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

[20] 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 Luce-style random utility, 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.

[21] 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 $b \to 0$.

[22] 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.

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, leading this system to put much more weight on distant past returns than the model-based system does, 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.

The pattern in Figure 1 holds robustly in our framework. We demonstrate this both numerically and analytically. In our numerical analysis, we consider 150 different parameterizations of our framework. Each parameterization corresponds to a set of parameter values $\{\alpha, \beta, \gamma, w\}$, where the values of the four parameters are drawn from $\alpha \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, $\beta \in \{10, 30, 50\}$, $\gamma \in \{0.3, 0.6, 0.9, 0.97, 0.99\}$, and $w \in \{0, 1\}$, and where $\alpha$ determines the values of all four learning rates $\alpha_{\pm}^{MF}$ and $\alpha_{\pm}^{MB}$. Half of these parameterizations correspond to model-free learning ($w = 0$) and half to model-based learning ($w = 1$). The remaining parameters are set to $L = T = 30$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$. For each of the 150 parameterizations, we repeat the analysis in Figure 1: we take 300,000 investors who make decisions according to our framework, regress their time 30 allocations on the past 30 years of stock market returns, and record the 30 coefficients. To see if these coefficients are positive on average, we compute their mean value; and to see if they are declining, we fit them to the exponential function $Ce^{-\delta(T-t)}$, where $C$ is a scaling factor, $t$ is the year a particular coefficient corresponds to, and $\delta > 0$ signifies coefficients that decline, the further back we go in time.

We find that, among the 75 model-based parameterizations with $w = 1$, all 75 exhibit coefficients with a positive mean that decline the further back we go in time. The same holds for 73 of the 75 model-free parameterizations with $w = 0$. Very robustly, then, the allocations recommended by the model-free and model-based systems put weights on past stock market returns that are positive and decline as we go further into the past. It is also robustly true that the decline in the coefficients is much more pronounced for the model-based system. Across the 75 model-based parameterizations, the average value of $\delta$ is 0.79, which is much higher than the average value of $\delta$ across the 75 model-free parameterizations, namely 0.07. Moreover, if, for each model-based parameterization $\{\alpha, \beta, \gamma, w = 1\}$, we match it to the analogous model-free parameterization $\{\alpha, \beta, \gamma, w = 0\}$, then, in all 75 cases, the estimated

$\delta$ for the model-based parameterization is higher than the estimated $\delta$ for the model-free parameterization.

We also confirm the robustness of the pattern in Figure 1 analytically. In Section 5.3 and Internet Appendix F, in a simplified version of our framework, we prove that, for both the model-free and model-based systems, the allocations put weights on past stock market returns that are positive and decline as we move further into the past, and that the decline is much faster in the case of the model-based system.

What is the intuition for the pattern in Figure 1? First, we explain why the allocations recommended by the model-free and model-based systems depend positively on past returns. The answer is clear in the case of the model-based system. Following a good stock market return, an investor's perceived distribution of market returns assigns a higher probability to good returns and a lower probability to bad returns. This raises the model-based $Q$ values of all stock market allocations, but particularly those of high allocations, making it more likely that the investor will choose a high allocation going forward.

The intuition in the case of the model-free system is quite different. If the investor chooses a 20% stock market allocation and the market posts a high return, this "reinforces" the choice of 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 makes the investor want to persist with a low allocation; in the other, it makes him want to persist with 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, and hence more reinforcement, 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 going forward.

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 $\alpha_+^{MB}$ and $\alpha_-^{MB}$, 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, 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.[23] 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 $\alpha_+^{MF}$ and $\alpha_-^{MF}$, 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.

A common assumption in psychology-based models of investor behavior is that some investors have extrapolative demand: their demand for a financial asset depends positively on the asset's past returns, and especially so on its recent past returns.[24] The results in this section show that each of the model-based and model-free systems can provide a foundation for extrapolative demand. The model-based system does so in a way that is similar to that of existing models, in particular, models of extrapolative beliefs: following a sufficiently good stock market return, the investor perceives a higher expected market return going forward,

[23] For an example, consider the upper panel of Table A1 in Internet Appendix C. 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.

[24] A partial list of papers that study extrapolative demand, either theoretically or empirically, is Cutler, Poterba, and Summers (1990), De Long et al. (1990), Barberis and Shleifer (2003), Barberis et al. (2015, 2018), Cassella and Gulen (2018), Chen, Liang, and Shi (2022), Jin and Sui (2022), Liao, Peng, and Zhu (2022), Bastianello and Fontanier (2025), and Pan, Su, Wang, and Yu (2025).

and so chooses a higher allocation to the market.

The model-free system also provides a foundation for extrapolative demand – the allocation it proposes typically puts positive and declining weights on past returns – and does so in a way that is new to the finance literature. While it is common to think of extrapolative demand as stemming purely from beliefs, the model-free mechanism shows that it need have nothing to do with beliefs: beliefs about future outcomes play no role in model-free learning. Going further, 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 structure in Sections 4.1 and 4.2 to shed light on some puzzling disconnects between allocations and beliefs.

We end this section with some additional comparative statics that show the rich implications of model-free learning. 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 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. Earlier in this section, we saw that, for a wide range of parameterizations, the model-free allocation puts more weight on recent returns than on distant past returns. Figure 3 shows that, for a few parameterizations – those with a high degree of generalization or a low degree of exploration – the opposite can be true: the model-free allocation can put more weight on distant than on recent past returns. We explain the full intuition for the patterns in Figure 3 in Internet Appendix D.[25]

# 4 Applications

In this section, we show that our framework can shed light on a number of important empirical facts in finance. This is striking, for two reasons. First, in prior research, this framework has

[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.

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. To study the various applications, we start with the setup of Section 3. There are again $L+T+1$ dates, $t=-L,\dots,-1$, 0, $1,\dots$, $T$. Relative to Section 3, we make two modifications to make the framework more realistic. First, we allow for dispersion in learning rates across investors. Second, 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 each application, we conduct simulations in which each investor interacts with a different return sequence from time $t=-L$ to time $t=T$.

In Sections 4.1 to 4.5, we discuss five applications of our framework. To convey the idea behind an application, we will often start by illustrating it for a specific parameterization, which will remain fixed throughout Section 4. In this parameterization, each of the 300,000 investors 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 $\alpha_+^{MF}$, $\alpha_-^{MF}$, $\alpha_+^{MB}$, and $\alpha_-^{MB}$ independently from a uniform distribution with mean $\bar{\alpha}$ and width $\Delta$. The specific parameter values are $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$, so that investors put equal weight on the model-free and model-based systems.

Importantly, for each of the five applications we discuss, after first illustrating it for the above specific parameterization, we then show that it emerges robustly from our framework – in other words, that it holds for a wide range of parameterizations. To do this, we consider 600 different parameterizations of our framework, each one corresponding to a set of parameter values $\{\bar{\alpha}, \Delta, \beta, b, w\}$, where the value of each parameter is drawn from $\bar{\alpha} \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, $\Delta \in \{0, 0.4\}$, $\beta \in \{10, 30, 50\}$, $b \in \{0, 0.0577, 0.115, 0.23\}$, and $w \in \{0, 0.25, 0.5, 0.75, 1\}$. The remaining, less pivotal parameters are kept fixed at the values $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. We then compute the fraction of these 600 parameterizations for which we observe the application in question. Note that each of the five values of $w$ – recall that $w$ is the weight on the model-based system – corresponds to 120 parameterizations. To show that a particular application emerges more strongly when the model-free system plays a larger role, we will also compute, for each value of $w$, the fraction of the 120 parameterizations associated with it for which the application holds.

## 4.1 Allocations and beliefs: The frequency disconnect

Our framework can help to resolve two puzzling disconnects between investor beliefs and investor actions – one in the frequency domain, which we discuss in this section, and one in the cross-section of investors, which we address in the next section. We account for these puzzles by way of a deep property of our framework, namely that, of the two systems, only the model-based system has an explicit role for beliefs. The model-free system, by contrast, has no notion of beliefs: it does not construct a probability distribution over 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. For individual investors, and for some financial professionals, *beliefs* about future stock market returns depend heavily on recent past returns: following a year or two of high returns, these investors expect high returns in the future as well (Greenwood and Shleifer, 2014). By contrast, investor *allocations* to the stock market depend to a substantial extent even on distant past returns (Malmendier and Nagel, 2011).[26]

How does our framework capture this disconnect? 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 gives a response that puts heavy weight on recent returns: in Section 3, we saw that, with a constant learning rate, the beliefs generated by the model-based system are strongly influenced by recent returns. However, the investor's allocation is determined by both the model-based and model-free systems, and, as shown in Section 3, the model-free system puts substantially more weight on distant past returns than the model-based system does. In this way, our framework creates a wedge between actions and beliefs: allocations depend more heavily on distant past returns than beliefs do.

Figure 4 illustrates these points. For the specific parameterization described above, 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. This solid line 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, we are now allowing for multiple cohorts and for dispersion across investors in their learning rates. The multiple cohorts in particular make the solid line in Figure 4 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.

The dashed line in Figure 4 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.

We now show that this frequency disconnect is a robust implication of our framework. For

[26] We formalize this in the following way. Malmendier and Nagel (2011) use the normalized weights $(n+1-j)^{\lambda}/A$ to characterize the relationship between the allocation of an investor with $n$ years of experience and the return he experienced $j$ years earlier. They obtain an estimate of $\lambda \approx 1.3$. Suppose that we now take the same functional form 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 far more weight on recent returns.

each of the 600 parameterizations described at the start of Section 4, we repeat the exercise in Figure 4. We take 300,000 investors in six cohorts who make their decisions at each time according to the hybrid system. We regress their allocations at time 30 on the past 30 years of stock market returns each investor was exposed to, and fit the 30 coefficients to the functional form $C_a e^{-\delta_a(T-t)}$. We also take investors' time-30 beliefs about the future stock market return, regress them on the past 30 years of stock market returns, and fit the 30 coefficients to the functional form $C_b e^{-\delta_b(T-t)}$. We define a frequency disconnect as $\delta_a < \delta_b$, so that beliefs depend more heavily on recent returns than allocations do.

Recall that, of the 600 parameterizations, 480 put at least some weight on the model-free system, so that $w < 1$. We find that $\delta_a < \delta_b$ for every single one of these 480 parameterizations. As such, the frequency disconnect is a very robust prediction of our framework when the model-free system is engaged. The top-left graph in Figure 5 plots, for each of the five values of $w$ we consider, the average value of $\delta_b - \delta_a$ across the 120 parameterizations corresponding to that $w$. The figure shows that the frequency disconnect becomes larger as $w$ falls, in other words, as the investor puts more weight on the model-free system.

## 4.2 Insensitivity of allocations to beliefs

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 (2024), and Yang (2025).

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: when the learning rate is constant, the beliefs generated by the model-based system put 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 now examine this effect quantitatively. For each of the 600 parameterizations described at the start of Section 4, we take 300,000 investors and estimate the sensitivity of their allocations to their beliefs by running a regression of their time-30 allocations on the return they expect over the next year at time 30. The top-right graph in Figure 5 reports, for each of the five values of $w$ we consider, the average sensitivity across the 120 parameterizations corresponding to that value of $w$. The graph shows that the sensitivity decreases markedly as we lower $w$, in other words, as the model-free system plays a larger role; when $w = 1$, the sensitivity is more than five times higher than when $w = 0$. Moreover, for $w = 0.5$, the framework produces an average sensitivity close to that estimated by Giglio et al. (2021).[27]

## 4.3 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

[27] The top-right graph in Figure 5 shows that, even when the model-based system alone determines allocations ($w = 1$), the framework predicts a relatively low sensitivity of allocations to beliefs. This is because the investors are limited to allocations between 0% and 100%, even though, under model-based learning, their extrapolative beliefs often lead them to want to take allocations outside this range. This mechanism is not driving the low sensitivity generated by the model-free system: only a small fraction of investors choose boundary allocations under model-free learning. Neither is exploration the source of the low model-free sensitivity. Rather, the low sensitivity under model-free learning is due to the fact that beliefs play no role in model-free choices.

$t + 1$, $R_{m,t+1}$ – 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

$$(n+1-j)^{\lambda}/A, \qquad j = 1, 2, \ldots, n, \tag{25}$$

where $\lambda$ 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 the specific parameterization described at the start of Section 4, and 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 (25). The dashed line plots the functional form in (25) for the cohort in question, and the vertical dotted line marks the point at which the cohort enters financial markets.

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 allocations on past market returns as we move from a return these investors experienced – the return 15 years in the past – to one they did not experience, the return 16 years in the past. The kink is driven by investors' model-free system, which puts substantial weight even on a return experienced 15 years in the past, but no weight at all on returns before that. The graph also shows that, within the subset of returns that these investors experience, their allocation puts greater weight on more recent past returns. Both the model-free and model-based systems contribute to this pattern, although the model-based system does so more.

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

We now confirm that experience effects are a robust feature of our framework, in that we observe them for a wide range of parameterizations. For each of the 600 parameterizations described at the start of Section 4, we repeat the exercise in Figure 6: for the 50,000 investors in each cohort, we regress their time-30 allocations on the past 30 years of stock market returns, obtaining coefficients $\{c_t^{(k)}\}_{t=1}^{T}$ for cohort $k \in \{1, \ldots, 6\}$, and then check whether we observe an experience effect. To make this precise, note that cohort $k$ enters financial markets at time $\tau(k) = 5(k-1)$. By experience effect, we mean: (i) that for $k = 1, \ldots, 6$, the coefficients $\{c_t^{(k)}\}_{t=\tau(k)+1}^{T}$ are on average positive and decline as we go further into the past, in the sense that, if we fit the coefficients to the functional form $C_k \exp(-\delta^{(k)}(T-t))$, then $\delta^{(k)} > 0$; and (ii) that for $k = 2, \ldots, 6$, $c_{\tau(k)+1}^{(k)} > c_{\tau(k)}^{(k)}$ and $c_{\tau(k)+1}^{(k)} - c_{\tau(k)}^{(k)} > c_{\tau(k)+2}^{(k)} - c_{\tau(k)+1}^{(k)}$, so that the coefficients jump up at the time of entry more than they do in the period immediately after entry, thereby creating kinks like those in Figure 6.

The middle-left graph in Figure 5 plots, for each of the five values of $w$ we consider, the fraction of the 120 parameterizations corresponding to that $w$ for which we observe an experience effect. The figure shows that, for $w = 1$, when there is no model-free learning, none of the 120 parameterizations exhibits an experience effect. However, when $w < 1$, so that model-free learning plays a role, an experience effect is observed much more frequently; for example, when $w = 0.5$, this is almost always the case. When $w = 0$, so that there is only model-free learning, an experience effect is less frequent than when $w = 0.5$. The reason is that, as shown in Figure 3, for a few parameterizations, the model-free allocation can depend more on distant past returns than on recent returns; as such, condition (i) above is sometimes violated.[28]

[28] Model-free learning is one possible psychological foundation for experience effects, but there are others; see Malmendier and Wachter (2024) for a discussion.

## 4.4 Inertia

There is substantial inertia in households' allocations to the stock market over time (Agnew, Balduzzi, and Sunden, 2003; Ameriks and Zeldes, 2004). This inertia is often attributed to transaction costs, procrastination, or inattention.

In this section, we show that model-free learning offers a new way of thinking about inertia in investor holdings: specifically, we show that the model-free system leads to much greater inertia than the model-based system. To demonstrate this, we take the 600 parameterizations described at the start of Section 4. For each parameterization, we compute a simple measure of inertia, namely, the fraction of the 300,000 investors whose allocation at time 30 is the same as their allocation at time 29. The middle-right graph in Figure 5 plots the average value of this fraction across the 120 parameterizations that correspond to each of the five values of $w$ that we consider.

The figure shows that, as $w$ falls, so that model-free learning plays a larger role, the degree of inertia goes up dramatically. When $w = 1$, so that investors use only model-based learning, there is little inertia: investors stick with the same allocation slightly more than 10% of the time. Since they have extrapolative beliefs, their perceived expected return on the stock market shifts from one period to the next, leading to a shift in allocations. When $w = 0$, so that investors use only model-free learning, the measure of inertia is approximately 40%, more than three times higher.

The intuition for the greater inertia produced by the model-free system is that, over time, as an investor tries different allocations, there is a good chance that one of these will be highly rewarded. The investor is then likely to stick with that allocation going forward. This intuition is general: in any setting, as an individual tries different actions, one may be rewarded with a very positive outcome; the individual is then likely to stick with that action. As such, model-free learning may help to explain inertia in many settings, not just the one we consider here.

## 4.5 Dispersion in allocations

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.

In this section, we show that model-free learning offers a new way of thinking about dispersion in allocations. Specifically, we show that model-free learning leads to a similar level of dispersion as model-based learning. This is striking because the dispersion induced by model-free learning cannot be attributed to differences in beliefs or objective functions: the model-free system has no notion of beliefs, and all investors have the same objective function in (7).

To explore this, we take the 600 parameterizations described at the start of Section 4 and, for each one, as a measure of dispersion, compute the cross-sectional standard deviation of the 300,000 investors' stock market allocations at time 30. The bottom-left graph in Figure 5 shows, for each of the five values of $w$ that we consider, the average level of dispersion across the 120 parameterizations that correspond to that value of $w$.

The graph shows that, when $w = 1$, so that model-based learning alone guides allocations, there is substantial dispersion in allocations. This is primarily due to differences in beliefs, although probabilistic choice also plays a role. Remarkably, when $w = 0$, so that model-free learning alone drives choices, we observe a similar level of dispersion.

What drives the dispersion in the case of model-free learning, if not beliefs or objective functions? It is 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.

# 5 Additional Analyses

In this section, we discuss several additional pieces of analysis.

## 5.1 Rational benchmarks

Our implementation of model-free and model-based learning in Sections 3 and 4 assumes that each investor uses learning rates that are constant over time. This implies that neither system is fully rational: for the model-free $Q$ values to converge to the correct $Q^*$ in (12), a declining model-free learning rate is needed, as in (24); similarly, for the model-based $Q$ values to converge to the correct $Q^*$, the declining model-based learning rate in (23) is needed. We use constant learning rates for psychological realism: most of the psychology research that we draw on uses constant learning rates.

In Internet Appendix E, we examine what happens when we use more rational versions of model-free and model-based learning that feature declining learning rates. Specifically, we repeat all of the main analyses in Sections 3 and 4 for two cases.

In the first case, investors use both rational model-free learning and rational model-based learning. We find that this framework does a poor job capturing the empirical facts. Most important, in this framework, investor beliefs put equal weight on past stock market returns, in sharp contrast to survey data where household beliefs depend heavily on recent returns. This, in turn, means that the framework cannot capture the frequency disconnect and does a poor job matching experience effects: it is unable to explain why, within the set of returns an investor has experienced, his allocation puts more weight on recent returns.

We then consider the case where investors use rational model-free learning in combination with the benchmark model-based learning with constant learning rates. We find that this framework performs fairly well: while the quantitative match to the data is not as good as for the implementation in Section 4, it can nonetheless capture experience effects, a frequency disconnect, insensitivity of allocations to beliefs, inertia, and dispersion in allocations. This is a striking finding: it shows that the results in Sections 3 and 4 do not hinge on constant model-free learning rates, but follow even from a more rational version of model-free learning.

## 5.2 Predictions

The model-free component of our framework makes several predictions, but all of them trace back to one core prediction, namely that, today, an investor is more likely to take an action that, in his past experience, was rewarded in the period after he took it. In this section, we study ways of formulating this prediction so that, given suitable data, it can be tested.

If the model-free system is a significant driver of investor behavior, a good predictor of the investor's allocation at time $t$ will be $\arg\max_a Q_t^{MF}(a)$, the allocation with the highest estimated model-free $Q$ value at time $t$. In the extreme case where the investor uses only the model-free system, so that $w = 0$, and there is no exploration, this variable will perfectly predict the investor's allocation.

The difficulty is that, even with data on an investor's holdings and returns, we do not observe $Q_t^{MF}(a)$, as computed in (13), because we do not know the investor's values of $\alpha_{\pm}^{MF}$ and $\gamma$. We therefore explore ways of approximating $Q_t^{MF}(a)$ so that a test can be implemented.

One approach is to construct, for each investor at each time, the 11 quantities $\{e_t^{(1)}(a)\}$, $a = 0\%, 10\%, \ldots, 100\%$, where

$$e_t^{(1)}(a) = \sum_{s=1}^{t} 1_{a_{s-1}=a} \log R_{p,s} \tag{26}$$

if the allocation $a$ has been tried at least once before time $t$, and $e_t^{(1)}(a) = 0$ otherwise. For each allocation $a$, this quantity is the sum of the log portfolio returns in the periods after the investor tried allocation $a$. The logic is that, if, after trying an allocation $a$, the investor experiences a high log portfolio return, this will increase the model-free $Q$ value for that allocation. As such, $a_t^{(1),*} = \arg\max_a e_t^{(1)}(a)$ should be a good predictor of the investor's allocation at time $t$.

An alternative approach is to approximate the updating equation (13) more directly. For each investor, we can define

$$e_t^{(2)}(a) = (1-\alpha)e_{t-1}^{(2)} + \alpha(\log R_{p,t} + \gamma \max_{a'} e_{t-1}^{(2)}(a')) \tag{27}$$

if the investor took the action $a$ at time $t-1$, and $e_t^{(2)}(a) = e_{t-1}^{(2)}(a)$ otherwise. If we knew the investor's values of $\alpha$ and $\gamma$, then, generalization aside, $e^{(2)}(a)$ would equal $Q^{MF}(a)$. In reality, we do not know these values. However, one idea is to choose reasonable values of $\alpha$ and $\gamma$, implement (27) using these values regardless of the investor's actual values of $\alpha$ and $\gamma$, and then check if the resulting $a_t^{(2),*} = \arg\max_a e_t^{(2)}(a)$ is a good predictor of time-$t$ allocations.

We now examine, in simulated data, whether $a_t^{(1),*}$ and $a_t^{(2),*}$ are indeed good predictors of allocations. We consider 375 different parameterizations of our framework, each one corresponding to a set of parameter values $\{\bar{\alpha}, \beta, \gamma, w\}$, where the values are selected from $\bar{\alpha} \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, $\beta \in \{10, 30, 50\}$, $\gamma \in \{0.3, 0.6, 0.9, 0.97, 0.99\}$, and $w \in \{0, 0.25, 0.5, 0.75, 1\}$. The remaining, less pivotal, parameters are set to $L = T = 30$, $\Delta = 0$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$. For each parameterization, we take 300,000 investors in a single cohort that enters financial markets at time 0, and, for each investor at each time, we compute $e_t^{(1)}(\cdot)$ and $e_t^{(2)}(\cdot)$, where for $e_t^{(2)}$, we set $\alpha = 0.5$ and $\gamma = 0.97$ regardless of investors' actual values of these variables. We then run two univariate regressions: one of investors' time-30 allocations on $a_{30}^{(1),*} = \arg\max_a e_{30}^{(1)}(a)$, and one of investors' time-30 allocations on $a_{30}^{(2),*} = \arg\max_a e_{30}^{(2)}(a)$. We record the $R$-squared of each regression. After going through all 375 parameterizations, we compute, for each of the two predictors, the average $R$-squared across the 75 parameterizations corresponding to each of the five possible values of $w$.

The bottom-right graph in Figure 5 plots these average $R$-squared. The graph confirms that both $a_{30}^{(1),*}$ and $a_{30}^{(2),*}$ have the desired properties. When $w = 1$, so that the model-free system plays no role, the variables have essentially no predictive power for allocations. However, when $w = 0$, so that the model-free system has exclusive control, the two variables have significant predictive power. As such, the ability of $a_{30}^{(1),*}$ and $a_{30}^{(2),*}$ to predict allocations is diagnostic of whether model-free learning is actually influencing investor allocations.

The statement "allocations can be predicted by $a_{30}^{(1),*}$ and $a_{30}^{(2),*}$" goes well beyond the statement, derived from research on experience effects, that "allocations can be predicted by investors' experienced returns." Since our framework takes a stand on the source of experience effects, it offers a more structured prediction. Specifically, in our framework, it matters not only what returns an investor has experienced, but also what the investor's allocation was when the returns were experienced. For example, according to traditional experience effects, an investor who experiences a high stock market return is predicted to then choose a high allocation to the stock market. In our setting, if it happens that, at the time of the high stock market return, the investor had a low allocation to the market, he is then predicted to be comfortable sticking with that low allocation. The predictor variables $a_{30}^{(1),*}$ and $a_{30}^{(2),*}$ capture this mechanism.

To test whether, in reality, model-free learning influences investor behavior, we could in principle check whether $a_t^{(1),*}$ and $a_t^{(2),*}$ predict investors' actual allocations. The data requirements here are steep: we would need a dataset that tracks a large number of specific individuals over many years and records their allocations and returns at each time. If such data become available, the test is implementable.

### 5.3 Analytical results

It is challenging to prove analytical results about the properties of our framework. The main difficulty is that, in contrast to model-based approaches, the model-free system primarily updates only one $Q$ value in each period, that of the most recently-chosen action.

Nonetheless, we *are* able to prove some theoretical results, and these go well beyond anything available in prior research. In a much simplified version of our framework, we prove that the stock market allocation recommended by each of the model-free and model-based systems puts weights on past stock market returns that are positive and that decline as we go further into the past, and that this decline is more pronounced for the model-based system. This provides an analytical foundation for the simulation-based findings in Section 3. In the two corollaries that follow, we extend these results to directly address two of our key applications in Section 4. We prove that model-free learning leads to a lower sensitivity of allocations to beliefs than does model-based learning. And we prove that model-free learning generates a frequency disconnect, in that the investor's beliefs put greater relative weight on recent returns than his allocations do; model-based learning, by contrast, does not generate a frequency disconnect.

While we prove the theorems below in a simplified version of our framework, this has an important advantage: it shows that the applications we discuss in Sections 3 and 4 follow from the essential feature of our framework – that, after an action is taken, its value is updated based on the subsequent reward. Precisely because the auxiliary features of our framework are excluded from the simplified setting of our theorems, we can conclude that these other features are not crucial for the applications in Sections 3 and 4.

Our two main theorems, which we prove in Internet Appendix F, are:

**Theorem (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 $R_{m,s} \equiv R$ for all periods $s \geq 1$. 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 $\alpha$.

Given these assumptions, the following result holds:

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

**Theorem (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 $R_{m,s} \equiv R$ for all periods $s \geq 1$.

Given these assumptions,

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

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

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,1}} = \frac{\beta R^{\beta-1}}{(R^\beta+1)^2}(1-\alpha)^{t-1}. \tag{30}$$

Expressions (28) and (29) confirm that the allocations recommended by each of the model-free and model-based systems put positive and declining weights on past returns – both expressions decline monotonically as $k$ increases – and that the decline is more pronounced for the model-based system: the model-free coefficient in (28) is lower than the model-based coefficient in (29) for low values of $k$, but higher than the model-based coefficient for high values of $k$.

The following corollaries address two of our applications from Section 4 – the sensitivity of allocations to beliefs, and the frequency disconnect.

**Corollary (Insensitivity):** The same assumptions apply as in the above theorems. Under model-free learning and as $t \to \infty$, the sensitivity of allocations to beliefs is

$$\frac{\partial \mathbb{E}[a_t]}{\partial \mathbb{E}_t^p(R_{m,t+1})} \equiv \frac{\partial \mathbb{E}[a_t]/\partial R_{m,t}}{\partial \mathbb{E}_t^p[R_{m,t+1}]/\partial R_{m,t}} = \frac{\beta R^{2\beta-1}}{(R^\beta+1)^3}. \tag{31}$$

Under model-based learning, the sensitivity of allocations to beliefs is

$$\frac{\partial \mathbb{E}[a_t]}{\partial \mathbb{E}_t^p(R_{m,t+1})} \equiv \frac{\partial \mathbb{E}[a_t]/\partial R_{m,t}}{\partial \mathbb{E}_t^p[R_{m,t+1}]/\partial R_{m,t}} = \frac{\beta R^{\beta-1}}{(R^\beta+1)^2}. \tag{32}$$

For any $R \geq 0$ and $\beta > 0$, the model-free sensitivity measure in (31) is strictly smaller than the model-based sensitivity measure in (32).

**Corollary (frequency disconnect):** The same assumptions apply as in the above theorems. Under model-free learning and as $t \to \infty$, there exists a $k^*$ such that, for $0 \leq k < k^*$,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} < \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}},$$

and for $k > k^*$,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} > \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}}.$$

As such, there is a frequency disconnect: the investor's beliefs put greater relative weight on recent past market returns than his allocation does. Under model-based learning,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} \Big/ \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}} = \frac{\beta R^{\beta-1}}{(R^\beta + 1)^2} \tag{33}$$

is a constant independent of $k$. There is therefore no frequency disconnect.

## 5.4 Completeness and restrictiveness

The framework of Section 2 is able to match a number of empirical facts. However, it also has several parameters. This raises the concern that the framework matches the empirical facts because it is too flexible.

Recently, Fudenberg et al. (2022) and Fudenberg, Gao, and Liang (2025) propose a way of evaluating this concern by computing a model's "completeness" and "restrictiveness." Completeness measures how well the model matches the actual data, while restrictiveness measures how well it matches a typical simulated dataset. A low level of restrictiveness means that the model matches the typical simulated dataset quite well – a negative quality, as it indicates that, due to the model's flexibility, it can "explain anything." Ideally, then, a framework should have high levels of both completeness and restrictiveness.

We implement a calculation in the spirit of Fudenberg et al. (2022) and Fudenberg, Gao, and Liang (2025). We describe it in detail in Internet Appendix G and summarize it here. Among the 600 parameterizations of our framework described at the start of Section 4, we search for the one for which the framework best matches the actual data, as measured by the sum of squared errors (SSE). Here, the actual data is a 184-element vector of numbers that summarize the empirical dependence of allocations on past returns, the relationship between beliefs and past returns, and the sensitivity of allocations to beliefs; we define it precisely in the Appendix. We record the SSE for this best match and label it the framework's completeness. We then simulate 100,000 artificial datasets, where each artificial dataset is a 184-element vector of simulated values for the relationship between allocations, beliefs, and returns; we detail its construction in the Appendix. For each simulated dataset, we find the parameterization of our framework that best matches it and record the SSE. After doing this for all 100,000 simulated datasets, we record the average best-fit SSE across them and label it the framework's restrictiveness.

We find that our framework's best-fit SSE for the actual data is 0.115. Meanwhile, its average best-fit SSE across the 100,000 simulated datasets is 2.897, a much higher number. This is an encouraging result. It shows that, while the framework of Section 2 is able to match the actual data, it is much less able to match simulated data. Put differently, the framework not only has a high level of completeness; it also has a high level of restrictiveness – despite having several parameters, it is not so flexible as to be able to "explain anything."

## 5.5 Alternative action spaces

In Sections 3 and 4, we focused on one set of possible actions: 11 percentage allocations to the stock market, $\{0\%, 10\%, \ldots, 100\%\}$. In Internet Appendix H, we repeat the main analyses in Sections 3 and 4 for two alternative action spaces to see if, and how, our results change. In a traditional model-based setting, the choice of action space does not affect the investor's behavior; in a setting with model-free learning, it may.

In the first alternative action space, investors choose the number of shares of the stock market they want to invest in. If, at time $t$, an investor's wealth is $W_t$ and the stock market price is $P_t$, then the action space at that time ranges from 0 shares to $100\lfloor W_t/(100P_t)\rfloor$ shares in increments of 100 shares.

In the second alternative action space, actions are defined relative to the previous period's allocation: either "keep the same allocation as before," "increase the prior allocation by 10%," or "decrease the prior allocation by 10%." To analyze this, we need to introduce a state variable $s_t$, as in the original formulation in (3), namely the allocation in the prior period. The reason is that whether an investor wants to increase or lower his allocation is likely to depend strongly on whether his prior allocation was low or high.

We repeat the main analyses in Sections 3 and 4 for these two alternative action spaces. We find that, for the first alternative, the results are very similar to those presented in Sections 3 and 4. For the second alternative, the results differ quantitatively from those in Sections 3 and 4, but only to a modest degree; moreover, the qualitative patterns are the same. Overall, we view the implications and applications of Sections 3 and 4 as robust to using these alternative action spaces.

## 5.6 Comparison with models of inattention

One of the properties of model-free learning is that it generates inertia in investor allocations. It is therefore natural to compare our framework to another framework that is often used to think about inertia in allocations, namely one based on investor inattention.

We consider three models of inattention. All three take the model-based component of our framework, discard the model-free component, and instead introduce a form of inattention.

In the first approach, each investor updates his beliefs about stock market returns at each date as in equations (17)-(18). With probability $p$, he is attentive, and also makes an active adjustment to his portfolio allocation: he computes the model-based $Q$ values in (19)-(20) and then chooses an action probabilistically. However, with probability $1-p$, he is not attentive, and his allocation drifts passively.

In our second approach, the investor again updates his beliefs in each period according to equations (17)-(18). Moreover, in each period, he updates the model-based $Q$ values of all allocations, as in (19)-(20). Finally, in each period, he checks whether the expected $Q$ value of his new allocation, if he did make an active choice, exceeds the $Q$ value of his previously-chosen allocation by more than some transaction cost $c$. If this condition is satisfied, the investor chooses an action probabilistically, according to current $Q$ values. Otherwise, his allocation drifts passively.

Both of the above inattention models assume that the investor can effortlessly update his beliefs in each period. In reality, however, the investor may find it just as effortful to update his beliefs as to change his allocation. We therefore consider a third model, a variant of the first, in which, at each time, the investor is inattentive with probability $1-p$ and updates neither his beliefs nor his allocation; and with probability $p$, he updates his beliefs and model-based $Q$ values based on all the returns realized since his last belief update and then chooses an action probabilistically based on the $Q$ values.

We analyze all three of the above inattention models in detail: we study their predictions for the dependence of stock market allocations on past stock market returns; experience effects; the frequency disconnect; the sensitivity of allocations to beliefs; and inertia. We present the full results in Internet Appendix J and summarize them here.

The first two inattention models lead to similar conclusions. On some dimensions, they make similar predictions to the framework of Section 2: for high levels of inattention, they predict that allocations will depend significantly even on distant past returns; and they can generate a frequency disconnect, insensitivity of allocations to beliefs, and inertia.

Interestingly, though, they make a rather different prediction about experience effects, namely that, if an investor enters financial markets at time $t$, his allocation at time $T$ will typically put *more* weight on the most recent return he did not experience, $R_{m,t}$, than on the first return he did experience, $R_{m,t+1}$. The reason is that, when an investor enters financial markets, he is paying attention, and so takes account of the return just before he enters, $R_{m,t}$. However, one year later, he may not be paying attention and may therefore not account for the return at that time, $R_{m,t+1}$. By contrast, our Section 2 framework makes the opposite prediction, one that is more in line with the evidence on experience effects, namely that the investor will put more weight on $R_{m,t+1}$ than on $R_{m,t}$.

The third model of inattention in which investors are slow to update both beliefs and allocations makes similar predictions to the first two inattention models on some dimensions – it, too, has trouble generating realistic experience effects. However, it differs from our framework in an additional important way: it is less able to generate insensitivity of allocations to beliefs; specifically, it predicts a sensitivity more than double that of the first inattention model.

Overall, our analysis shows that, on some dimensions, our framework makes similar predictions to inattention models. Nonetheless, the two types of frameworks also differ significantly in some of their implications – most notably, regarding experience effects and the sensitivity of allocations to beliefs.

## 5.7 Other analyses

**Parameter estimation.** In Section 4, we showed that, for a wide range of parameterizations, our framework can account for a number of empirical facts about investor behavior. In Internet Appendix K, we describe a complementary analysis. We estimate the values of four key parameters – the mean model-based learning rate across investors, $\bar{\alpha}^{MB}$; the mean model-free learning rate $\bar{\alpha}^{MF}$; the exploration parameter $\beta$; and the weight $w$ on the model-based system – by searching for the values that best match 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.

Our estimates are $\bar{\alpha}^{MB} = 0.33$, $\bar{\alpha}^{MF} = 0.26$, $\beta = 20$, and $w = 0.38$, so that investors put substantial weight on both the model-free and model-based systems. This estimate is consistent with Figure 5, which shows that a value of $w$ between 0.25 and 0.5 successfully generates both experience effects and a sensitivity of allocations to beliefs similar to that in Giglio et al. (2021).

**State dependence and system performance.** 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 L, we show how state dependence can be introduced into our framework. Specifically, we allow for mean-reversion in returns, so that a state variable based on past returns has predictive power for future returns. We study the investment performance of the model-free and model-based systems in this setting, and find that they have similar performance. This points to one additional reason why some households might use model-free learning: the performance of their model-free system can be at least as good as that of their unsophisticated model-based system.

**Other model-free systems.** The two model-free algorithms most commonly used by psychologists to model human behavior are Q-learning and SARSA. In this paper, we have used Q-learning as our model-free algorithm. In Internet Appendix M, we show that, if we repeat our main analyses with SARSA, we obtain similar results. This supports the claim that our results do not depend on the precise form of model-free learning. Rather, they follow from the essential feature of this learning, one that is common to all model-free algorithms, namely that, at each time, the value of the most-recently action is updated based on the outcome it led to.

# 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 empirical facts about investor allocations and beliefs. Through the model-based system, our framework preserves a role for beliefs in driving 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

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 (2025), "Expectations and Learning from Prices," *Review of Economic Studies* 92, 1341-1374.

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

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

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

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 (2024), "Insensitive Investors," *Journal of Finance* 79, 2473-2503.

Charness, G. and D. 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.

Enke, B., and T. Graeber (2023), "Cognitive Uncertainty," *Quarterly Journal of Economics* 138, 2021-2067.

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.

Fudenberg, D., Gao, W., and A. Liang (2025), “How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories,” *Review of Economics and Statistics,* forthcoming.

Fudenberg, D., Kleinberg, J., Liang, A., and S. Mullainathan (2022), “Measuring the Completeness of Economic Models,” *Journal of Political Economy* 130, 956-990.

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

Gabaix, X. (2019), “Behavioral Inattention,” *Handbook of Behavioral Economics,* North Holland.

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

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

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

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

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

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

Lattimore, T. and C. Szepesvari (2020), *Bandit Algorithms*, Cambridge University Press.

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 (2024), “Memory of Past Experiences and Economic Decisions,” *Oxford Handbook of Human Memory,* 2228-2266.

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 (2025), "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*, Springer Nature.

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.

Yang, J. (2025), "On the Decision-Relevance of Subjective Beliefs," Working paper.

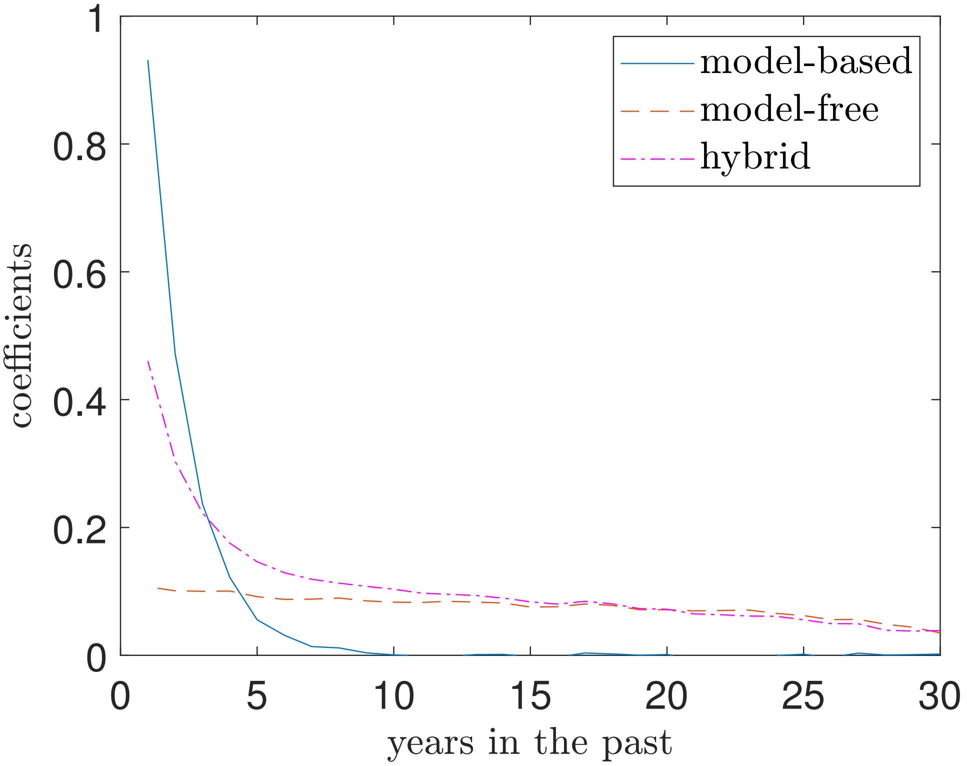

**Figure 1.** We run a regression of investors' allocations to the stock market $a_T$ at time $T$ on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system.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

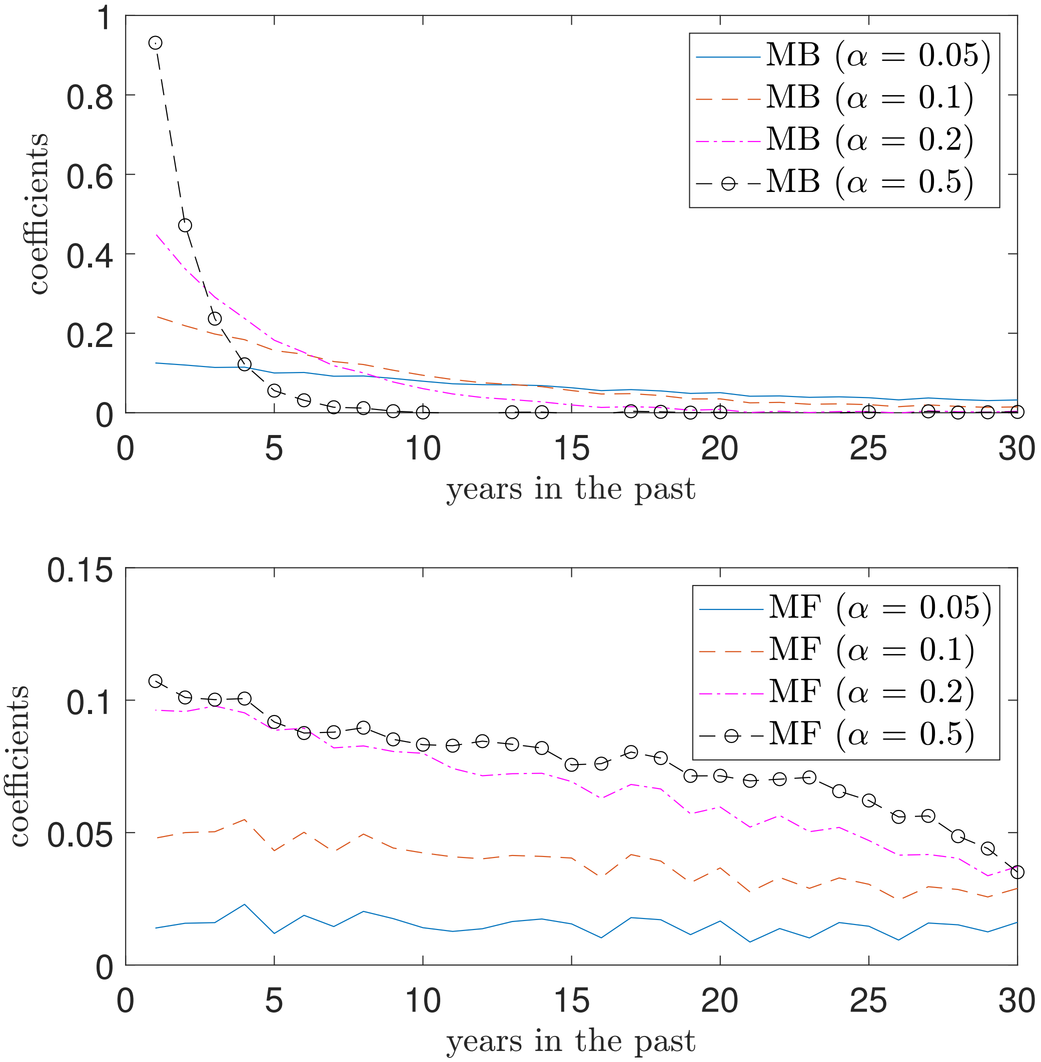

**Figure 2.** We run a regression of investors' allocations to the stock market $a_T$ at time $T$ on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to. The top graph plots the coefficients for the model-based system for four values of the learning rates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

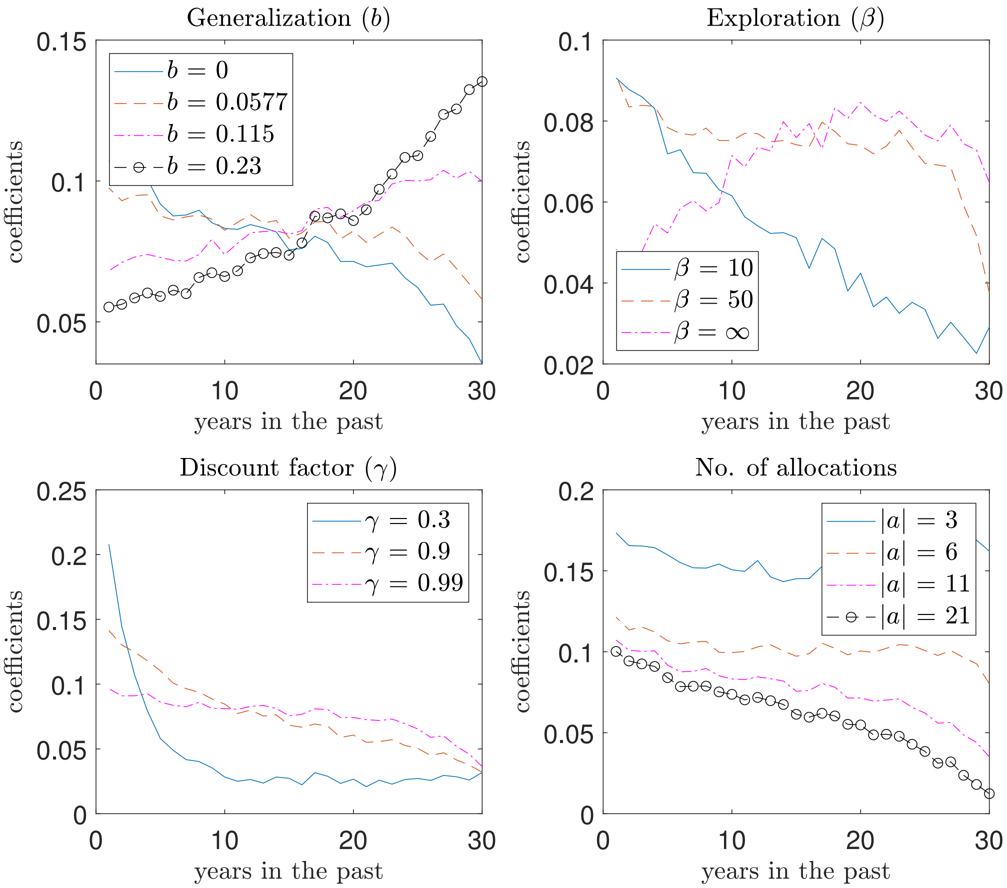

**Figure 3.** For different sets of parameter values, we run a regression of investors' allocations to the stock market $a_T$ at time $T$ under the model-free system on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients. The lines 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 $\beta$, namely 10, 50, and $\infty$, which corresponds to no exploration; to three values of the discount factor $\gamma$, 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.

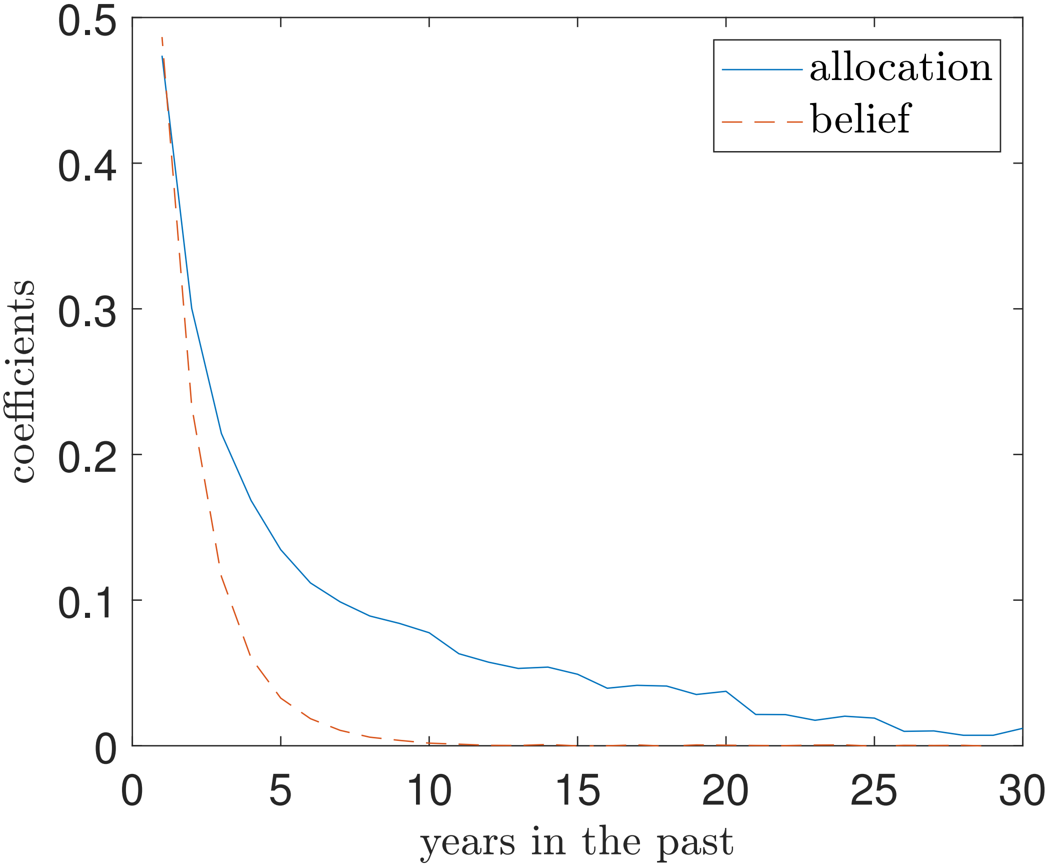

**Figure 4.** The solid line plots the coefficients in a regression of the stock market allocation $a_T$ at date $T$ chosen by investors who use a hybrid system to make decisions on the past 30 years of stock market returns the investors were exposed to. The dashed line plots the coefficients in a regression 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 $\alpha_{+}^{MF}$, $\alpha_{-}^{MF}$, $\alpha_{+}^{MB}$, and $\alpha_{-}^{MB}$ is drawn independently from a uniform distribution with mean $\bar{\alpha}$ and width $\Delta$. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0.0577$, and $w = 0.5$.

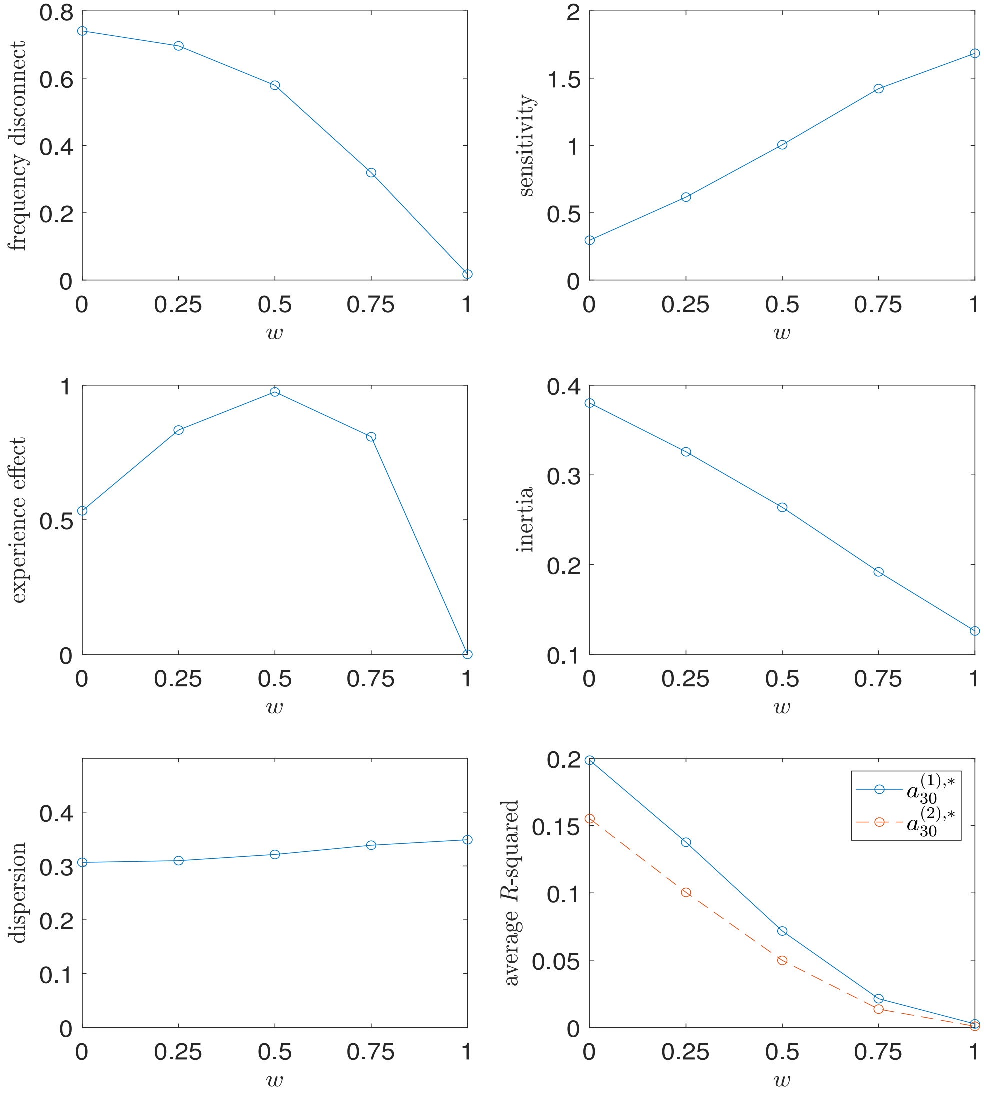

**Figure 5.** For each of a large number of parameterizations of our framework, we check whether the framework generates: a frequency disconnect; insensitivity of allocations to beliefs; an experience effect; inertia in allocations; dispersion in allocations; and predictability of allocations by the experience variables $a_{30}^{(1),*}$ and $a_{30}^{(2),*}$. For each of these six phenomena, the six graphs record, for each of the five values of $w$, namely 0, 0.25, 0.5, 0.75, and 1, where $w$ is the weight on the model-based system, the fraction of parameterizations with that value of $w$ for which the phenomenon is observed.

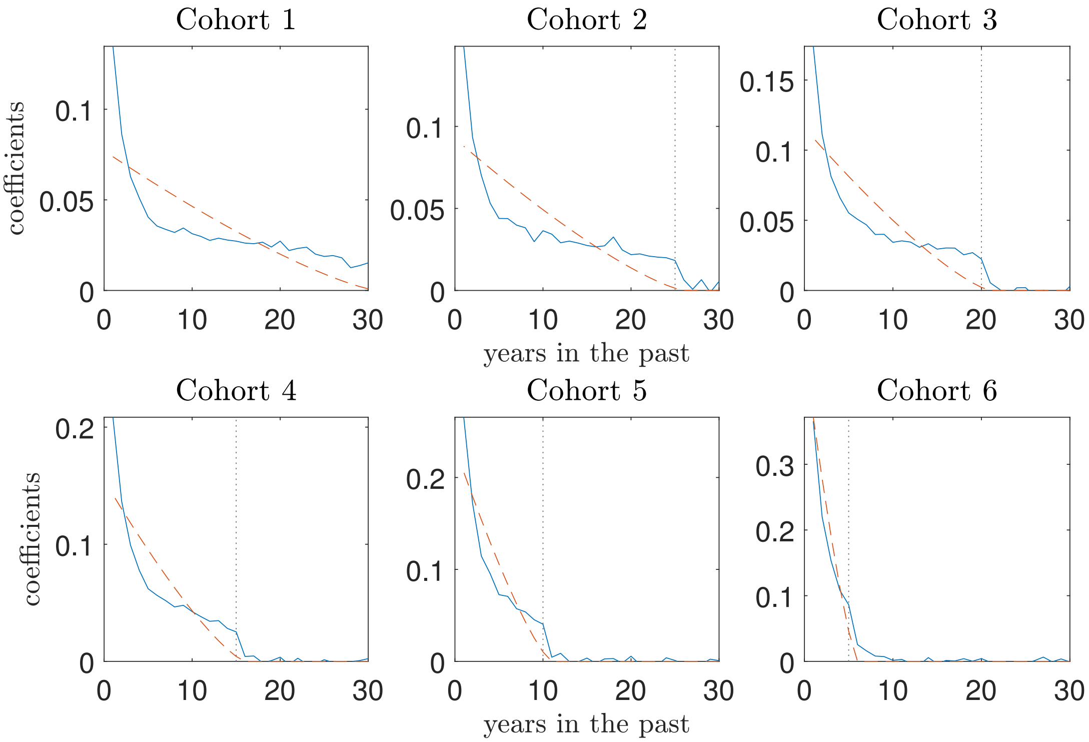

**Figure 6.** The six graphs correspond to six cohorts of investors. In each graph, the solid line plots the coefficients, normalized to sum to one, in a regression of the time-$T$ stock market allocations $a_T$ of the investors in that cohort on the past 30 years of stock market returns they were exposed to. The six cohorts have different numbers of years of experience, namely $n = 30$, 25, 20, 15, 10, and 5; 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 $\alpha_+^{MF}$, $\alpha_-^{MF}$, $\alpha_+^{MB}$, and $\alpha_-^{MB}$ is drawn independently from a uniform distribution with mean $\bar{\alpha}$ and width $\Delta$. 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.

# INTERNET APPENDIX

## A. Experimental Evidence of Model-free Decision-making

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]

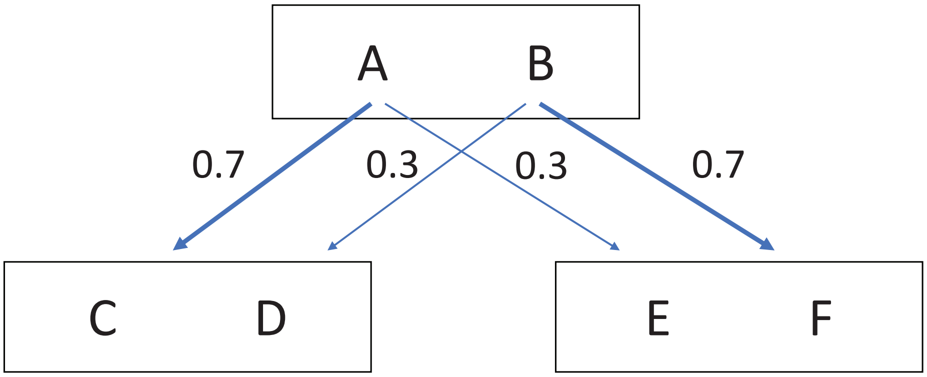

**Figure A1.** The diagram shows the structure of an experiment in Daw et al. (2011). In the first stage, the participant has a choice between two options, A and B; in the second stage, he chooses either between options C and D or between options E and F. The arrows indicate the transition probabilities 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% weight on the model-based system.[2]

The above experiment illustrates a tension between model-free and model-based learning. 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. A similar 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. Analysis of Alternative Decision Problem Formulations

Consider a setting with two risky assets – risky asset $A$ and risky asset $B$ – and define the "market" asset, denoted as asset $M$, as the sum of one share of asset $A$ and one share of asset $B$. In this section, we examine whether the behavior of an investor who uses model-free learning depends on whether his control variables are his allocations to assets $A$ and $B$; his allocations to assets $A$ and $M$; or his allocations to assets $B$ and $M$.

We assume that assets $A$ and $B$ have lognormal returns:

$$\begin{aligned} \log R_{A,t} &= \mu_A + \sigma_A \cdot \varepsilon_{A,t}, \\ \log R_{B,t} &= \mu_B + \sigma_B \cdot \varepsilon_{B,t}, \end{aligned} \tag{1}$$

where, for simplicity, we take $\varepsilon_{A,t}$ and $\varepsilon_{B,t}$ to be two mutually independent standard Normal

[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.

random variables. At each point in time, asset $M$ consists of a 50% allocation to asset $A$ and a 50% allocation to asset $B$ – at the end of each period, there is rebalancing between asset $A$ and asset $B$ so that, moving forward, asset $M$ continues to consist of a 50% allocation to asset $A$ and a 50% allocation to asset $B$. Formally, for any $t$, the return on asset $M$ is given by

$$R_{M,t+1} = 0.5R_{A,t+1} + 0.5R_{B,t+1}. \tag{2}$$

For each investor, at each point in time, his action space consists of a pair of portfolio weights. In the case where the investor is choosing allocations to assets $A$ and $B$, these are given by

$$\mathbf{a}_t = (a_{A,t}, a_{B,t}), \tag{3}$$

where $a_{A,t}$ and $a_{B,t}$ are each chosen from the 11 possible allocations $\{0\%, 10\%, \ldots, 100\%\}$. We require that $a_{A,t} + a_{B,t} \leq 1$; that is, the investor is not allowed to borrow. So, for each value of $a_{A,t}$, the feasible range of $a_{B,t}$ goes from 0% to $1 - a_{A,t}$.

We choose the asset parameters $(\mu_A, \sigma_A, \mu_B, \sigma_B)$ so that the optimal allocations $a^*_A$ and $a^*_B$ satisfy $0 \leq a^*_A, a^*_B \leq 1$ and $0 \leq a^*_A + a^*_B \leq 1$. Here, we set $\mu_A = 0\%$, $\sigma_A = 20\%$, $\mu_B = -1\%$, and $\sigma_B = 20\%$. We also set $R_f$ to 1. The optimal allocations are then given by

$$(a^*_A, a^*_B) = \arg\max_{a_A, a_B} \mathbb{E}\left[\log\left((1 - a_A - a_B)R_f + a_A R_{A,t+1} + a_B R_{B,t+1}\right)\right]. \tag{4}$$

The numerical solutions are $a^*_A = 50\%$ and $a^*_B = 20\%$; the remaining 30% of the investor's wealth is allocated to the risk-free asset. The level of expected utility is 0.0062.

Below, we consider three different scenarios. The first scenario allows the investor to trade assets $A$ and $B$; the remaining wealth is allocated to the risk-free asset. The second scenario allows the investor to trade assets $A$ and $M$, and the third scenario allows him to trade assets $B$ and $M$. In a rational model-based environment, where the investor understands that asset $M$ is always an equal mix of assets $A$ and $B$, he can achieve his optimal portfolio in any of three ways: investing 50% of his wealth in asset $A$ and 20% in asset $B$; investing 30% of his wealth in asset $A$ and 40% in asset $M$; or investing −30% of his wealth in asset $B$ and 100% in asset $M$. The question is: What does a model-*free* investor do, in each of these three scenarios?

**Scenario 1.** We consider the most basic model-free learning algorithm, one with constant learning rates and without generalization. In Scenario 1, the investor trades asset $A$ and asset $B$. Note that here, the $Q$ values are two dimensional functions, as suggested by (3). Other than the four asset parameters ($\mu_A = 0\%$, $\sigma_A = 20\%$, $\mu_B = -1\%$, and $\sigma_B = 20\%$), the remaining parameters take the values listed in the caption to Figure 1 in the main text. As before, the fraction of wealth allocated to each asset is one of 11 possible percentage allocations. The sum of the two fractions must be less than or equal to 100%; in other words, the investor cannot borrow.

Using equation (1), we simulate 30 years of returns for assets $A$ and $B$. For this sequence of returns, we compute the investor's allocations to assets $A$ and $B$ at each date $t = 0, 1, \ldots, 30$. The left panel in Figure A2 plots these allocations. Below, we will compare these allocations to those chosen under the two alternative scenarios.

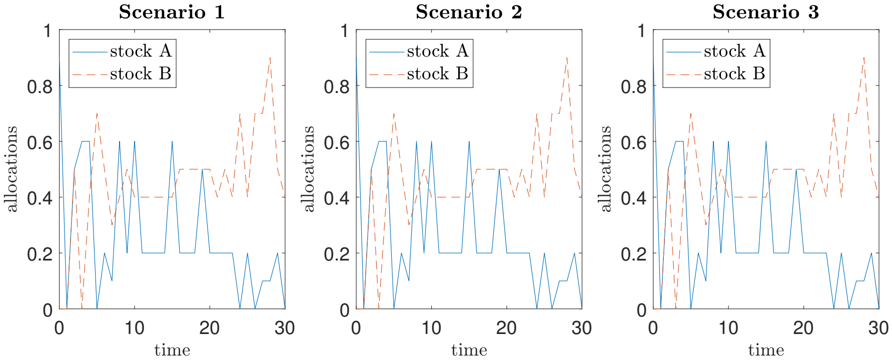

**Figure A2.** An investor uses model-free learning to allocate his portfolio in an economy with two risky assets, $A$ and $B$, and a market asset that is a 50:50 combination of $A$ and $B$. The left, middle, and right graphs plot the investor's effective allocations to assets $A$ and $B$ when his control variables are the allocations to assets $A$ and $B$, the allocations to assets $A$ and $M$, and the allocations to assets $B$ and $M$, respectively.

**Scenario 2.** In this scenario, the investor trades asset $A$ and asset $M$. Here, the investor's action space consists of a pair of portfolio weights:

$$\mathbf{a}_t = (a_{A,t}, a_{M,t}). \tag{5}$$

We assume that $a_{M,t}$ is chosen from the 11 possible allocations $\{0\%, 20\%, 40\%, \ldots, 200\%\}$. For a given level of $a_{M,t}$, the allocation $a_{A,t}$ ranges from $-\frac{1}{2}a_{M,t}$ to $1-a_{M,t}$, in 10% increments. This construction of the action space effectively allows the investor to choose the same set of allocations to asset $A$ and asset $B$ as in Scenario 1, through holding a combination of asset $A$ and asset $M$.

We now take the same simulated returns for assets $A$ and $B$ that we used in Scenario 1 and, for these returns, compute the investor's allocations to assets $A$ and $M$ at each date $t = 0, 1, \ldots, 30$. In this exercise, in order to focus purely on the impact of the action space, we keep the randomness associated with the stochastic choice the same as in Scenario 1. Once we have computed the allocations to assets $A$ and $M$, we calculate the effective allocations to assets $A$ and $B$ and plot these in the center panel of Figure A2. We note that, even though the control variables are different across Scenarios 1 and 2, the effective allocations to assets $A$ and $B$ turn out to be exactly identical.

**Scenario 3.** In this scenario, investors trade asset $B$ and asset $M$. The investor's action space consists of a pair of portfolio weights:

$$\mathbf{a}_t = (a_{B,t}, a_{M,t}). \tag{6}$$

We assume that $a_{M,t}$ is chosen from the 11 possible allocations $\{0\%, 20\%, 40\%, \ldots, 200\%\}$. For a given level of $a_{M,t}$, the allocation $a_{B,t}$ ranges from $-\frac{1}{2}a_{M,t}$ to $1-a_{M,t}$, in 10% increments. This construction of the action space effectively allows the investor to choose the same set of allocations to asset $A$ and asset $B$ as in Scenario 1, through holding a combination of asset

$B$ and asset $M$.

We now take the same simulated returns for assets $A$ and $B$ that we used in Scenarios 1 and 2 and, for these returns, compute the investor's allocations to assets $B$ and $M$ at each time $t = 0, 1, \ldots, 30$. As before, to focus purely on the impact of the action space, we keep the randomness associated with the stochastic choice the same as in Scenarios 1 and 2. Once we have computed the allocations to assets $B$ and $M$, we calculate the effective allocations to assets $A$ and $B$ and plot these in the right panel of Figure A2. We note that, even though the control variables in Scenario 3 are different from those in Scenarios 1 and 2, the effective allocations to assets $A$ and $B$ are exactly the same across all three scenarios. As such, the choice of action space has no impact on behavior.

To check that the result in Figure A2 is not a fluke, we repeat the exercise 1,000 times. In other words, we simulate 1,000 different 30-year sequences of returns on assets $A$ and $B$, and for each one, compute the investor's effective allocations to assets $A$ and $B$ under each of the three scenarios. We find that, for each of the 1,000 iterations, the effective allocations to $A$ and $B$ are identical at each moment of time across all three scenarios.

The main takeaway of the analysis in this section is that, under model-free learning, the three scenarios examined above all lead to the *same* portfolio allocations: in Scenarios 2 and 3, converting the holding of asset $M$ to holdings of assets $A$ and $B$ leads to identical holdings to those in Scenario 1. This finding makes sense: model-free investors are not concerned about holding specific assets *per se*; instead, they focus on learning how their actions yield rewards – namely, the log portfolio returns resulting from holding assets $A$ and $B$, assets $A$ and $M$, or assets $B$ and $M$.

## C. The Mechanics of the Model-free and Model-based Systems: An Example

In this section, we illustrate the mechanics of the model-free and model-based systems by way of an example. 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 $\sigma = 0.2$. We set the investor's learning rates to $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, the exploration parameter $\beta$ to 30, the discount factor $\gamma$ to 0.97, and the degree of generalization $b$ to 0.0577. At each date, we allow the investor to choose his stock market allocation $a_t$ from one of the 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$.

In the framework of Section 2, 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 consider two simpler cases in this section: one where the investor uses only the model-free system to make decisions, and one where he uses only the model-based system.

Table A1 shows the model-free $Q$ values, $Q^{MF}$, based on equations (11), (13), and (14) in the main text (upper panel) and the model-based $Q$ values, $Q^{MB}$, based on equations (19) and (20) in the main text (lower panel) that the investor assigns to the 11 possible allocations 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 panel indicates that the investor chose a 70% allocation at date 0.[3]

Consider the upper panel of Table A1. 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) in the main text, the algorithm also engages in some generalization: since a 60% allocation and an 80% allocation are similar to a 70% allocation, their $Q$ values also fall, albeit to a lesser extent. The $Q$ values of more distant allocations are unaffected, at least to three decimal places.

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

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 allocations. 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 A1 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] 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 expression (11) in the main text. 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.

**Table A1. 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 $L = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

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

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

The upper panel of Table A1 raises the concern that the model-free system generates too much "bouncing around" in allocations. In fact, in general, the model-free system does *not* generate a lot of bouncing around of allocations. As shown in Section 4.4 of the paper, the model-free system generates strong inertia in allocations, both in absolute terms and relative to the model-based system. The reason why, in the upper panel, the chosen allocation varies a lot, is that we went out of our way to select, from various simulated examples, one that featured a lot of variation in allocations, as this makes it easier to explain the mechanics of the model-free system.

By contrast, as can be seen in the lower panel of Table A1, the model-*based* system really does generate a lot of variation in allocations. This sets up the puzzle that needs to be explained. In survey data, household beliefs about future stock market returns are extrapolative: they depend positively on the past year or two of stock market returns. But as soon as we embed such beliefs in a portfolio choice framework, they predict large swings in portfolio allocations, contrary to the infrequent moves we observe in practice. The model-based system captures this puzzle: the model-based learners' process for constructing a distribution of stock market returns leads these investors to weight recent returns heavily in their beliefs; when coupled with their utility function, this leads to big swings in allocations, as seen in the lower panel of Table A1.

Again, these big swings set up the puzzle: If households have extrapolative beliefs, how can this nonetheless lead to inertia in allocations? Traditional finance offers one explanation, namely transaction costs or attention costs. Our paper offers a new possibility, namely model-free learning.

### D. 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 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, the model-free allocation starts to put 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 $\beta$, which controls the degree of exploration, namely 10, 50, and $\infty$. As $\beta$ 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 $\beta$ – any $\beta$ 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 $\beta$ 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 $\beta$ 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 $\gamma$, namely 0.3, 0.9, and 0.99. As we lower $\gamma$, 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.

### E. Analysis of Rational Benchmarks: Results

Our implementation of model-free and model-based learning in Sections 3 and 4 assumes that each investor uses learning rates that are constant over time. This implies that neither system is fully rational: for the model-free $Q$ values to converge to the correct $Q^*$ in equation (12) in the main text, a declining model-free learning rate is needed, as in (24) in the main text; similarly, for the model-based $Q$ values to converge to the correct $Q^*$, the declining model-based learning rate in (23) in the main text is needed. We use constant learning rates for the sake of psychological realism: most of the psychology research that we draw on uses constant learning rates.

In this section, we examine what happens when we use more rational versions of model-free and model-based learning that feature declining learning rates. We find that, for the applications in Section 4, it is important that the model-based system use a constant learning rate; without it, we cannot match some key facts about investor behavior. By contrast, the constant learning rate is *not* needed for the model-free system: even with a declining model-free learning rate, we continue to match the facts about investor behavior discussed in Section 4, although the quantitative fit is not quite as good. This is an example of the robustness of our results to the specific implementation of model-free learning: the results do not hinge on a constant model-free learning rate.

For the rational model-based system, we adopt the declining learning rate in equation (23) in the main text. Each investor then proceeds as before: he constructs his perceived return distribution as in (17)-(18) in the main text and his model-based $Q$ values as in (19). Consistent with the assumption that returns are i.i.d., the declining learning rate leads the investor to put equal weight on all past stock market returns when forming beliefs.

For the rational model-free system, we take inspiration from research on multi-armed bandit problems where, similar to our setting, an individual selects among different options by trying them and observing the outcome. Specifically, we use:

$$Q_t^{MF}(a) = Q_{t,1}^{MF}(a) + \frac{\gamma}{1-\gamma} \max_{a'} Q_{t,1}^{MF}(a') \tag{7}$$

and

$$Q_{t,1}^{MF}(a) = \frac{\sum_{k=0}^{t-1} 1_{a_k=a} \log R_{p,k+1}(a_k)}{\sum_{k=0}^{t-1} 1_{a_k=a}} \tag{8}$$

if the allocation $a$ has been tried at least once before time $t$, and $Q_{t,1}^{MF} = 0$ otherwise. Equation (7) has the same form as equation (12) in the main text; $Q_{t,1}^{MF}(a)$ is an estimate of $E(\log((1-a)R_f + aR_{m,t+1}))$, where the estimate is constructed as the average log portfolio return in the periods after taking allocation $a$. This approach effectively uses a declining learning rate – an action-specific learning rate that declines over time based on how often an action has been tried.

At each time, and for each action, the investor then computes a hybrid $Q$ value as in equation (21) in the main text. In one last modification, we assume that the investor chooses an action at each time not using the softmax approach in equation (22) in the main text, but rather using another device that is common in research on multi-armed bandits, namely an "epsilon-greedy" algorithm where, at time $t$, the investor takes the allocation with the highest estimated $Q^{HYB}$ value with probability $1-\varepsilon$, and with probability $\varepsilon$ chooses one of the other actions at random. This ensures that, in the limit as $t \to \infty$, the investor will try all actions infinitely often, which in turn means that each of $Q^{MF}$, $Q^{MB}$, and $Q^{HYB}$ will converge to the correct $Q^*$.[4]

[4] We use the epsilon-greedy algorithm for this exercise because convergence to $Q^*$ is assured without further assumptions. In the case of softmax, convergence to $Q^*$ is assured with the additional assumption

We now repeat the main analyses in Sections 3 and 4 for two cases: first, the case where investors use the rational versions of both model-based and model-free learning; and second, the case where they use the rational version of model-free learning and the benchmark version of model-based learning from Sections 3 and 4 that features a constant learning rate. (The results in the case where investors use the rational version of model-based learning and the benchmark version of model-free learning are very similar to the results in the case where they use rational versions of both algorithms.)

**Rational model-based and rational model-free systems.** Figure A3 is the analog of Figure 1: it shows how the allocations recommended by each of the model-free, model-based, and hybrid systems at time 30 depend on the past 30 years of stock market returns. To construct this graph, we take a single cohort of 300,000 investors, each of whom observes a different sequence of past stock market returns. We set $L = T = 30$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. In the case of the hybrid system, $w = 0.5$.

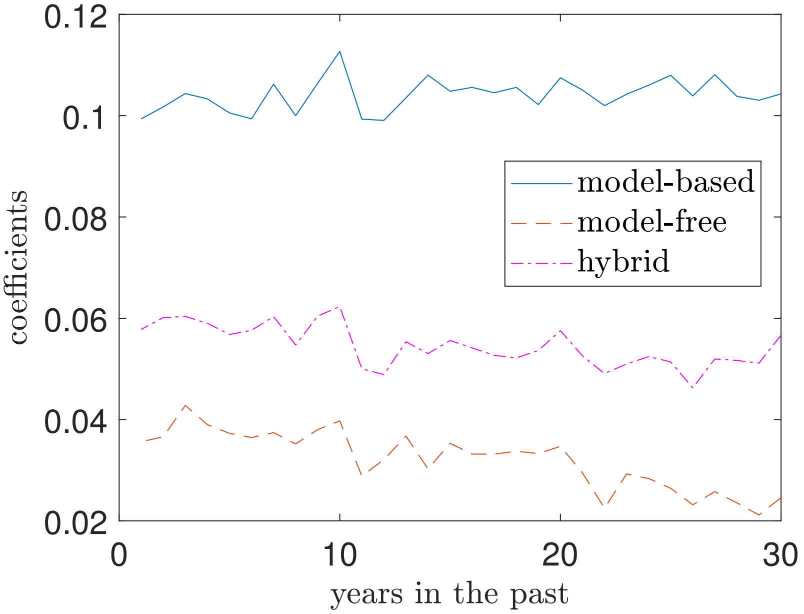

**Figure A3.** Analogous to Figure 1 in the main text, the graph shows how the allocations recommended by the model-free, model-based, and hybrid systems depend on past stock market returns. In contrast to Figure 1, investors use rational versions of the model-free and model-based systems. There are 300,000 investors. We set $L = T = 30$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. In the case of the hybrid system, $w = 0.5$.

The main difference between Figure A3 and Figure 1 is for the model-based system. In Figure 1, the line for this system declines sharply; in Figure A3, it is flat: since, under the rational model-based system, investor beliefs put equal weight on all past returns, the model-based allocation does too. By contrast, the model-free lines in Figure A3 and Figure

that all actions are tried infinitely often. In practice, the results we report in this section are similar, whether we use epsilon-greedy or softmax.

1 are similar – an early indication that our results are robust to using rational model-free learning.

Figure A4, the analog of Figure 4, presents the results for the frequency disconnect, while Figure A5, the analog of Figure 6, presents the results for experience effects. To construct these figures, we take 300,000 investors in six cohorts and set $L = T = 30$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $w = 0.5$. We also compute the sensitivity of allocations to beliefs; the measure of inertia from Section 4.4; and the measure of dispersion from Section 4.5. We find these to be, respectively, 2.52, 28.57%, and 31.27%.

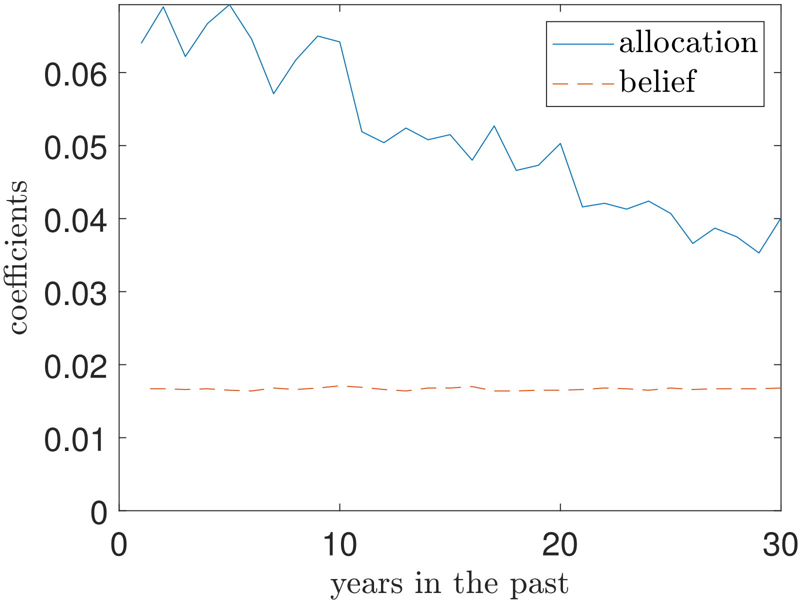

**Figure A4.** Analogous to Figure 4 in the main text, the graph shows how investors' allocations and beliefs at time 30 depend on the past 30 years of stock market returns. In contrast to Figure 4, investors use rational versions of the model-free and model-based systems. There are 300,000 investors in six cohorts. We set $L = T = 30$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $w = 0.5$.

These results show that a framework that combines rational model-based learning with rational model-free learning does a poor job capturing the empirical facts. Most important, in this framework, investor beliefs put equal weight on past stock market returns, in sharp contrast to survey data where household beliefs depend heavily on recent returns. This, in turn, means that the framework cannot capture the frequency disconnect and that it does a poor job matching experience effects: Figure A5 shows that, within the set of returns an investor has experienced, he does not put more weight on recent returns, which is counterfactual.

**Rational model-free and benchmark model-based systems.** We now consider the case where investors use the benchmark model-based system from Sections 3 and 4, one with a constant learning rate, together with the rational model-free system in equations (7)-(8).

Figure A6 is the analog of Figure 1 in the main text: it shows how the allocations recommended by the model-based, model-free, and hybrid systems at time 30 depend on the

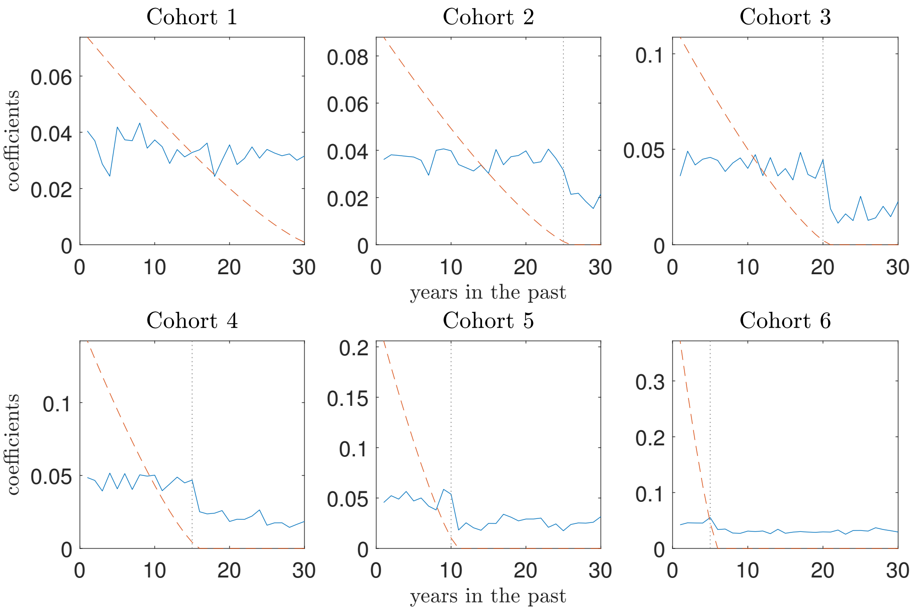

**Figure A5.** Analogous to Figure 6 in the main text, the graph shows how the allocations of each of the six cohorts depend on past stock market returns. In contrast to Figure 6, investors use rational versions of the model-free and model-based systems. There are 300,000 investors in six cohorts. We set $L = T = 30$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $w = 0.5$.

past 30 years of stock market returns. To construct this graph, we take a single cohort of 300,000 investors, each of whom observes a different sequence of past stock market returns. We set $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. In the case of the hybrid system, we set $w = 0.5$. We see that the graph is similar to Figure 1: replacing the baseline model-free system with a rational one leads to similar results.

Figure A7, the analog of Figure 4, presents results for the frequency disconnect, while Figure A8, the analog of Figure 6, presents results for experience effects. To construct these figures, we take 300,000 investors in six cohorts and set $L = T = 30$, $\varepsilon = 0.5$, $\alpha_{\pm}^{MB} = 0.5$, $\Delta = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $w = 0.5$. The sensitivity of allocations to beliefs is 0.9572; the measure of inertia is 24.03%; and the measure of dispersion is 34.44%.

These results convey an important finding, namely that our results in Sections 3 and 4, obtained with a model-free system with a constant learning rate, are robust to using a more rational model-free system with a declining learning rate: a framework with a rational model-free system can capture experience effects, a frequency disconnect, insensitivity of allocations to beliefs, inertia, and dispersion in allocations. Nonetheless, we use a constant learning rate in Sections 3 and 4 in order to be psychologically realistic: psychology research overwhelmingly uses a constant learning rate. And while a rational model-free system can qualitatively capture the applications we consider, the quantitative fit is not quite as good, particularly in the case of experience effects and the frequency disconnect.

## F. Analytical Results

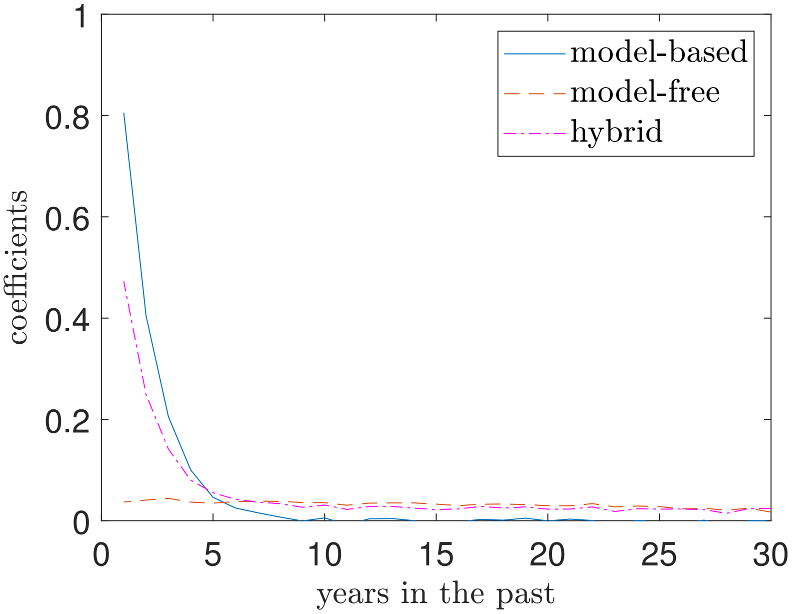

**Figure A6.** Analogous to Figure 1 in the main text, the graph shows how the allocations recommended by the model-free, model-based, and hybrid systems depend on past stock market returns. In contrast to Figure 1, investors use a rational version of the model-free system. There are 300,000 investors. We set $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. In the case of the hybrid system, $w = 0.5$.

In this Appendix, we prove the theorems stated in Section 5.3, which are labeled as Theorem 3 and Theorem 4 below, and also the two corollaries from that section, which speak directly to two of our key applications: the frequency disconnect and the sensitivity of allocations to beliefs. To build intuition for the proofs, we start with two simpler theorems, Theorem 1 and Theorem 2, which assume a learning rate of $\alpha = 1$ for both the model-free and model-based systems.

**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 $t > 0$,

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

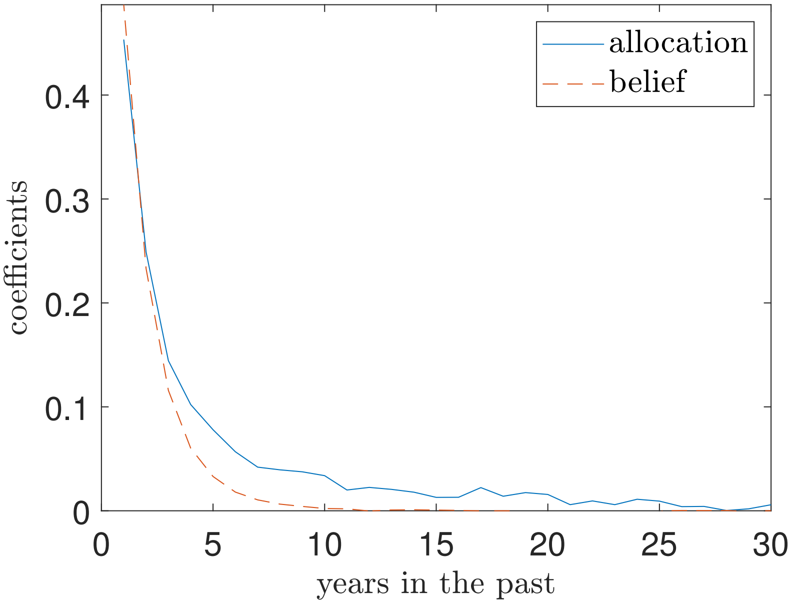

**Figure A7.** Analogous to Figure 4 in the main text, the graph shows how investors' allocations and beliefs at time 30 depend on the past 30 years of stock market returns. In contrast to Figure 4, investors use a rational version of the model-free system. There are 300,000 investors in six cohorts. We set $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $w = 0.5$.

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:

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

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,

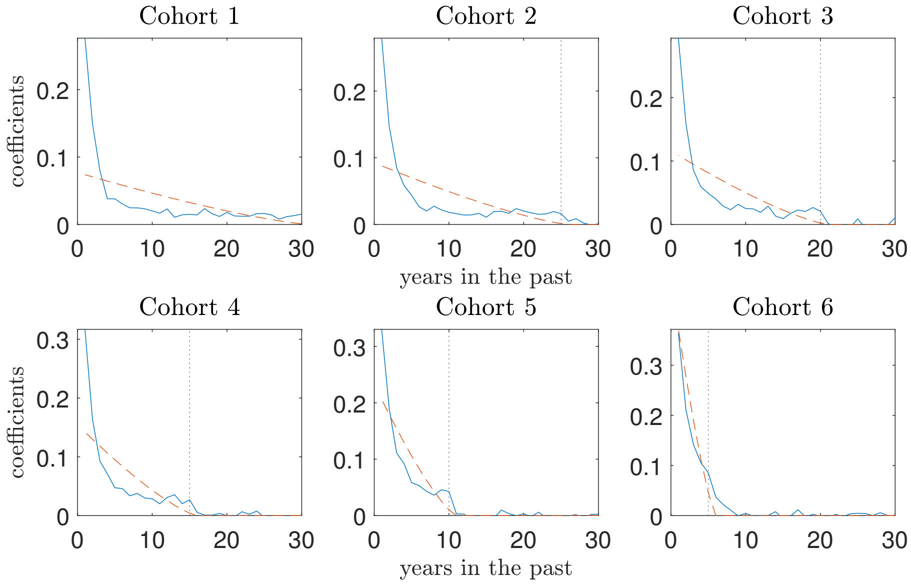

**Figure A8.** Analogous to Figure 6 in the main text, the graph shows how the allocations of each of the six cohorts depend on past stock market returns. In contrast to Figure 6, investors use a rational version of the model-free system. There are 300,000 investors in six cohorts. We set $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\varepsilon = 0.5$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $w = 0.5$.

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

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

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

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

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

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

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

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:

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

**Proof:** At any time $t > 0$,

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

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 $\alpha$.

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

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

We now turn to (24). We have

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

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

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

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. ■

**Corollary (Insensitivity):** The same assumptions from Theorems 3 and 4 apply. Under model-free learning and as $t \to \infty$, the sensitivity of allocations to beliefs is

$$\frac{\partial \mathbb{E}[a_t]}{\partial \mathbb{E}_t^p(R_{m,t+1})} \equiv \frac{\partial \mathbb{E}[a_t]/\partial R_{m,t}}{\partial \mathbb{E}_t^p[R_{m,t+1}]/\partial R_{m,t}} = \frac{\beta R^{2\beta-1}}{(R^\beta+1)^3}. \tag{40}$$

Under model-based learning, the sensitivity of allocations to beliefs is

$$\frac{\partial \mathbb{E}[a_t]}{\partial \mathbb{E}_t^p(R_{m,t+1})} \equiv \frac{\partial \mathbb{E}[a_t]/\partial R_{m,t}}{\partial \mathbb{E}_t^p[R_{m,t+1}]/\partial R_{m,t}} = \frac{\beta R^{\beta-1}}{(R^\beta+1)^2}. \tag{41}$$

For any $R \geq 0$ and $\beta > 0$, the model-free sensitivity measure in (40) is strictly smaller than the model-based sensitivity measure in (41).

**Proof:** The expected return is

$$\mathbb{E}_t^p(R_{m,t+1}) = (1-\alpha)^{t-1} R_{m,1} + \alpha \sum_{j=2}^{t} (1-\alpha)^{t-j} R_{m,j}.$$

As a result,

$$\frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}} = \begin{cases} \alpha(1-\alpha)^k & 0 \leq k \leq t-1 \\ (1-\alpha)^{t-1} & k = t-1 \end{cases}. \tag{42}$$

Combining equation (21) with (42) gives (40). Combining (34) and (35) with (42) gives (41). Moreover, the ratio of (40) and (41) is

$$\frac{R^\beta}{R^\beta+1} < 1 \tag{43}$$

for any $R \geq 0$ and $\beta > 0$. ■

**Corollary (frequency disconnect):** The same assumptions from Theorems 3 and 4 apply. Under model-free learning and as $t \to \infty$, there exists a $k^*$ such that, for $0 \leq k < k^*$,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} < \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}},$$

and for $k > k^*$,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} > \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}}.$$

Under model-based learning,

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} / \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}} = \frac{\beta R^{\beta-1}}{(R^\beta+1)^2} \tag{44}$$

is a constant independent of $k$.

**Proof:** For model-free learning, $t \to \infty$, and $k < t-1$, taking the ratio of (21) and (42) gives

$$\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}} / \frac{\partial \mathbb{E}_t^p(R_{m,t+1})}{\partial R_{m,t-k}} = \frac{\beta R^{2\beta-1}}{(R^\beta+1)^3} \left( \frac{R^\beta + 1 - \alpha R^\beta}{(R^\beta+1)(1-\alpha)} \right)^k \tag{45}$$

for model-free learning. Note that

$$\frac{R^\beta + 1 - \alpha R^\beta}{(R^\beta + 1)(1 - \alpha)} = \frac{R^\beta + 1 - \alpha R^\beta}{R^\beta + 1 - \alpha R^\beta - \alpha} > 1.$$

So the right-hand side of (45) is a monotonically increasing function of $k$ that goes to infinity as $k$ increases. When $k = 0$, this function equals

$$\frac{\beta R^{2\beta-1}}{(R^\beta + 1)^3} < \frac{1}{R^\beta + 1} < 1. \tag{46}$$

As a result, there exists a $k^*$ such that, for $0 \leq k < k^*$, the right-hand side of (45) is less than one, and for $k > k^*$, it is greater than one.

For model-based learning, taking the ratio of (34)-(35) and (42) gives (44). ■

## G. Computation of Completeness and Restrictiveness

In Section 5.4, we summarize an analysis of our framework's "completeness" and "restrictiveness," concepts put forward by Fudenberg et al. (2022) and Fudenberg, Gao, and Liang (2025). In this section, we provide the details of the analysis.

We work with the 600 parameterizations described at the start of Section 4. These are indexed by $\{\bar{\alpha}, \Delta, \beta, b, w\}$, where the values of the five parameters are drawn from the sets $\bar{\alpha} \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$, $\Delta \in \{0, 0.4\}$, $\beta \in \{10, 30, 50\}$, $b \in \{0, 0.0577, 0.115, 0.23\}$, and $w \in \{0, 0.25, 0.5, 0.75, 1\}$. The remaining parameters are set to $L = T = 30$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$. We consider an economy with six cohorts and 100,000 investors in each cohort. For each parameterization, we record the following 184 coefficients, which we label "model outputs":

- The first 30 coefficients are for cohort 1. They are the slope coefficients in a regression of time-30 allocations on stock market returns from the past 30 years. Coefficient 1 is for the stock market return from 30 years ago, while coefficient 30 is for the stock market return from last year.

- Coefficients 31 to 60, 61 to 90, 91 to 120, 121 to 150, and 151 to 180 are the analogs of coefficients 1 to 30 for cohorts 2, 3, 4, 5, and 6, respectively.

- Coefficient 181 is the coefficient in a regression of final allocations on final beliefs; here, we combine all six age cohorts.

- For coefficients 182 to 184, we run a regression of final beliefs on stock market returns over the past 30 years, combining all six age cohorts. Coefficient 182 corresponds to the stock market return from last year, coefficient 183 to the stock market return from two years ago, and coefficient 184 to the stock market return from three years ago.

Given these model outputs, we compute the completeness measure as follows. For each of the 600 parameterizations, we compare each of the first 180 recorded coefficients, normalized for each cohort, with its empirical target, namely the cohort-specific weight based on the

Malmendier and Nagel (2011) functional form. Specifically, for cohort $n$ and the stock market return from past year $k$, the weight is given by

$$w(k,n) = \begin{cases} \dfrac{(31-k-5(n-1))^{\lambda}}{\sum_{k'=1}^{30-5(n-1)} (31-k'-5(n-1))^{\lambda}} & k \leq 30-5(n-1) \\ 0 & k > 30-5(n-1) \end{cases}, \tag{47}$$

where $\lambda = 1.3$. We compare coefficient 181 with the empirical value of one, based on Giglio et al. (2021). For the coefficients in the regression of final beliefs on stock market returns, coefficient 182 is compared to the empirical value of 0.127; coefficient 183 is compared to the empirical value of 0.037; and coefficient 184 is compared to the empirical value of 0.029. To obtain these empirical values, 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.

For each parameterization, we compute the sum of squared errors across the 184 coefficients. The completeness measure is the smallest sum of squared errors among the 600 parameterizations. We find this to be 0.1145.

Next, we compute the restrictiveness measure. We create 100,000 simulated datasets. Each simulated dataset is a 184-element vector. For each simulated dataset, coefficients 1 to 180, which summarize the potential dependence of allocations on past returns, are computed as follows. First, we generate

$$\hat{w}(k,n) = \begin{cases} \dfrac{c_0^{1,n} + c_1^{1,n}(31-k) + c_2^{1,n}(31-k)^2}{\sum_{k'=1}^{30-5(n-1)} \left[c_0^{1,n} + c_1^{1,n}(31-k') + c_2^{1,n}(31-k')^2\right]} & k \leq 30-5(n-1) \\ \dfrac{c_0^{2,n} + c_1^{2,n}(31-k) + c_2^{2,n}(31-k)^2}{\sum_{k'=31-5(n-1)}^{30} \left[c_0^{2,n} + c_1^{2,n}(31-k') + c_2^{2,n}(31-k')^2\right]} & k > 30-5(n-1) \end{cases}, \tag{48}$$

where $c_0^{1,n}$ and $c_0^{2,n}$ are drawn from Unif$(-0.5, 0.5)$, $c_1^{1,n}$ and $c_1^{2,n}$ are drawn from Unif$(-0.5, 0.5)/30$, and $c_2^{1,n}$ and $c_2^{2,n}$ are drawn from Unif$(-0.5, 0.5)/900$; here, Unif$(a, b)$ denotes the uniform distribution between $a$ and $b$. Second, we bound $\hat{w}(k,n)$ from above by 1 and from below by $-1$. These bounded $\hat{w}(k,n)$ are the simulated coefficients 1 to 180. Intuitively, we are using two polynomials with random coefficients to generate the dependence of allocations on returns; the two polynomials correspond to the periods before and after the investors enter financial markets. Coefficient 181, which represents the potential sensitivity of allocations to beliefs, is randomly and uniformly drawn from the interval $(0, 4)$; coefficients 182 to 184, which represent the potential dependence of beliefs on past returns, are each randomly and uniformly drawn from the interval $(-0.2, 0.3)$.

For each of the 100,000 simulated datasets, we compute the minimum sum of squared errors across the 600 parameterizations, where the squared errors are derived by comparing the model outputs and the simulated coefficients described above. The restrictiveness measure is the average minimum sum of squared errors across the 100,000 simulated datasets. We find this to be 2.897.

## H. Alternative Action Spaces

In Sections 3 and 4 of the main text, we focus on a particular set of possible actions: 11 percentage allocations to the stock market, $\{0\%, 10\%, \ldots, 100\%.\}$ In Figure 3 of the main text, we consider finer and coarser sets of percentage allocations. However, there are other possible action spaces – for example, one where the investor chooses the number of shares of the stock market that he wants to hold; or one where actions are defined relative to the prior allocation, as in "choose an allocation 10% higher than before." In a traditional model-based framework, the choice of action space does not affect the investor's behavior. In a setting with model-free reinforcement learning, it may.

In this section, we study this issue. We repeat the main analyses in Sections 3 and 4 for the two alternative action spaces listed above. We find that, while there are some quantitative changes in our results, particularly for the second alternative, the results are nonetheless qualitatively similar. As such, we view the implications and applications of Sections 3 and 4 as being robust to using these alternative action spaces.

**Number of shares.** We start by studying the action space where the investor chooses the number of shares of the stock market that he wants to hold. We take the setting of Section 3, where the timeline runs from $t = -L$ to $t = T$ and where the model-based and model-free systems begin operating at $t = -L$ and $t = 0$, respectively. There are 300,000 investors. At time 0, they each have \$10,000 and the initial stock market price is \$10 per share. The action space at time 0 consists of 11 possible actions: {0 shares, 100 shares, 200 shares,. . . , 1000 shares}. At each subsequent date $t$, as the investor's wealth $W_t$ and stock market price $P_t$ vary, the action space also shifts, and ranges from 0 shares to $100\lfloor W_t/(100P_t)\rfloor$ shares in increments of 100 shares.

Figure A9, the analog of Figure 1 in the main text for this alternative action space, plots the dependence of the model-free, model-based, and hybrid allocations on past stock market returns; the parameter values are the same as for Figure 1. We note that Figure A9 closely resembles Figure 1.

Next, we revisit our applications from Section 4. To do so, we allow for six cohorts of investors; for dispersion across investors in the learning rates $\alpha_{\pm}^{MF}$ and $\alpha_{\pm}^{MB}$; and for generalization. Figure A10 presents the results for experience effects; it is the counterpart to Figure 6, and the two figures use the same parameter values. Figure A11, which is the analog of Figure 4, is for the frequency disconnect. Finally, for the benchmark parameter values, the sensitivity of allocations to beliefs is 1.167. We note that, for all three applications, the results for this alternative action space are very similar to the results in Sections 3 and 4 for the original action space.

In producing the above results, we have to make an assumption as to which action space the investor uses at time $t + 1$ to update the $Q$ values – and specifically, to compute $\max_{a'} Q_t^{MF}(a')$ and to implement generalization. For the results presented here, we assume that the action space is the one the investor selected from at time $t$. We have repeated our analysis for the case where, at time $t+1$, the investor updates using the action space he will select from at time $t + 1$. The results are almost identical.

**Relative action choice.** We now consider an alternative action space, one where the investor's possible actions are defined relative to the allocation chosen in the previous period.

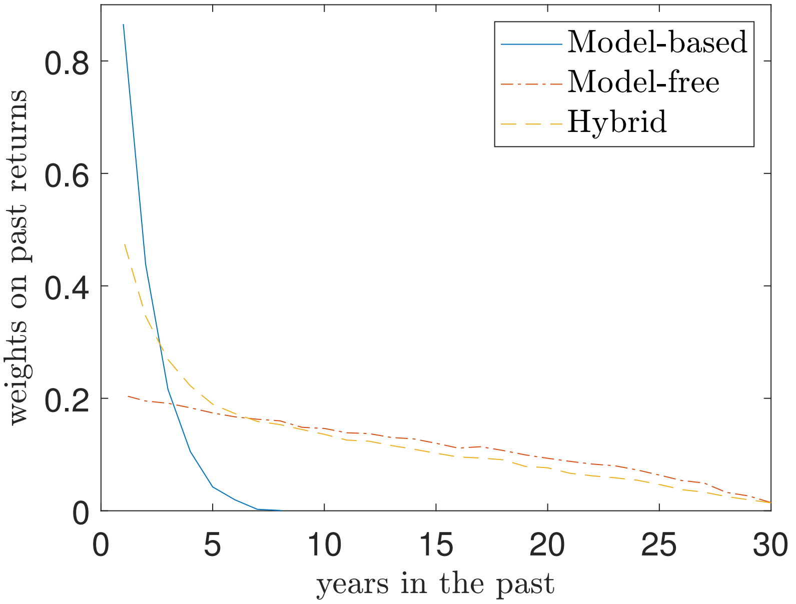

**Figure A9.** Analogous to Figure 1, the graph shows how the allocations recommended by the model-free, model-based, and hybrid systems depend on past stock market returns. In contrast to Figure 1, investors choose how many shares of the market to hold. There are 300,000 investors. We set $L = T = 30$, $\alpha^{MF}_{\pm} = \alpha^{MB}_{\pm} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$. In the case of the hybrid system, $w = 0.5$.

Specifically, there are three possible allocations: "keep the allocation the same as before," "choose an allocation that is 10% higher than the previous allocation," and "choose an allocation that is 10% lower than the previous allocation." We denote this action space as $\{-0.1, 0, 0.1\}$. In contrast to much of the analysis in the main text, we now introduce a state variable, namely the "pre-adjustment" stock market allocation – the percentage allocation to the stock market prior to any adjustment of –10%, 0%, or 10%; it is natural that, whether an investor wants to increase or decrease his allocation to the stock market should depend on whether his pre-adjustment allocation was high or low.

The model-free updating rule, in the absence of generalization, is

$$Q^{MF}_{t+1}(a_t, w_{t_-}) = Q^{MF}_t(a_t, w_{t_-})$$
$$+\alpha^{MF}_{t,\pm}\left(\log R_{p,t+1}(w_{t_-}+a_t) + \gamma \max_{a'} Q^{MF}_t(a', \underbrace{\tfrac{(w_{t_-}+a_t)R_{m,t+1}}{(w_{t_-}+a_t)R_{m,t+1}+(1-w_{t_-}-a_t)R_f}}_{w_{(t+1)_-}}) - Q^{MF}_t(a_t, w_{t_-})\right) \quad (49)$$

where $w_{t_-}$ is the time-$t$ portfolio weight the investor assigns to the stock market *prior to* making the adjustment based on action $a_t$.

The timing of the updating rule is as follows. At time $t_-$, the investor has the pre-adjustment portfolio weight $w_{t_-}$; this is the current fraction of his wealth invested in the stock market. He then chooses an action $a_t$. His post-adjustment portfolio weight at time $t$

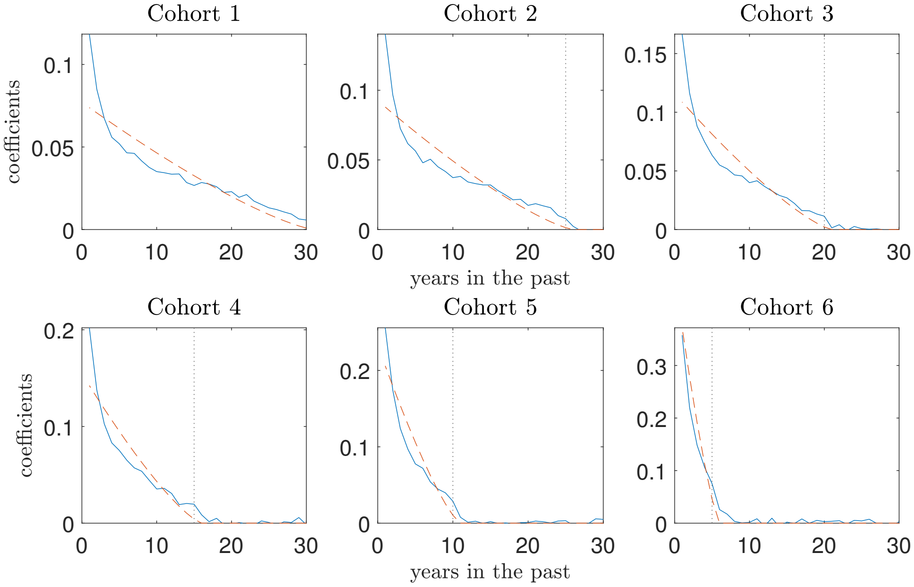

**Figure A10.** Analogous to Figure 6, the graph shows how the allocations of each of the six cohorts depend on past stock market returns. In contrast to Figure 6, investors choose how many shares of the market to hold. There are 300,000 investors. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0$, and $w = 0.5$.

is then $w_{t_-} + a_t$. If, at any time, $w_{t_-} + a_t$ is above 100%, it is capped at 100%; and if it is below 0%, it is floored at 0%. One period later, at time $(t+1)_-$, given the realized stock market return from $t$ to $t+1$, namely, $R_{m,t+1}$, the new pre-adjustment portfolio weight is

$$w_{(t+1)_-} = \frac{(w_{t_-} + a_t)R_{m,t+1}}{(w_{t_-} + a_t)R_{m,t+1} + (1 - w_{t_-} - a_t)R_f}. \tag{50}$$

The investor then chooses an action $a_{t+1}$, and the post-adjustment portfolio weight at time $t+1$ is $w_{(t+1)_-} + a_{t+1}$. This process repeats over time.

Despite the fact that action $a$ can take only three discrete values, namely –10%, 0%, or 10%, the portfolio weight $w_{t_-}$ takes continuous values because the stock market return itself takes continuous values. To make the above algorithm tractable – to avoid having a state variable that can take an infinite number of values – we make an approximation and have only a finite set of 11 states {0%, 10%, 20%, . . ., 100%}. The updating rule becomes

$$\begin{aligned} Q^{MF}_{t+1}(a_t, s_{t_-}) &= Q^{MF}_t(a_t, s_{t_-}) \\ +\alpha^{MF}_{t,\pm} &\left(\log R_{p,t+1}(w_{t_-} + a_t) + \gamma \max_{a'} Q^{MF}_t(a', s_{(t+1)_-}) - Q^{MF}_t(a_t, s_{t_-})\right), \end{aligned} \tag{51}$$

where $s_{t_-}$ is the state among the 11 possible states that is closest to $w_{t_-}$; that is, $|s_{t_-} - w_{t_-}| \leq$ 5%.

We now turn to model-based learning. Here, for all three possible values of action $a$, the

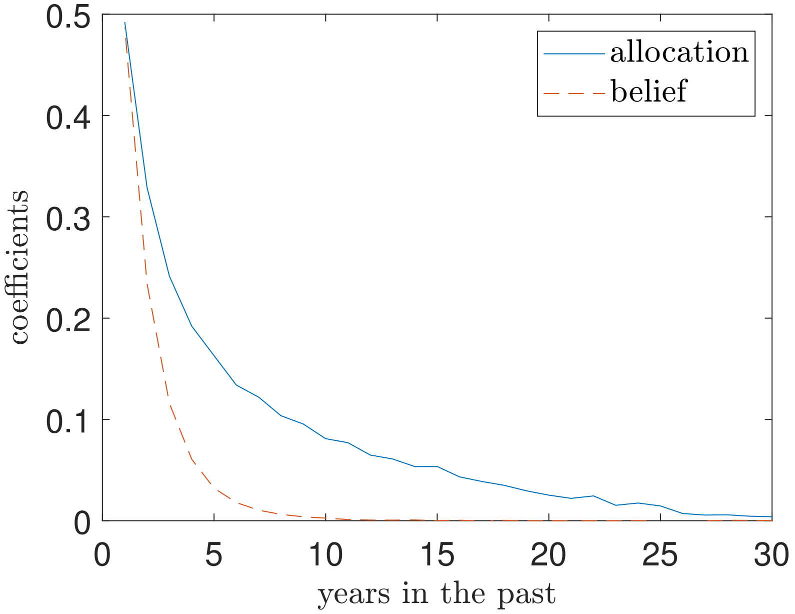

**Figure A11.** Analogous to Figure 4, the graph shows how investors' allocations and beliefs at time 30 depend on the past 30 years of stock market returns. In contrast to Figure 4, investors choose how many shares of the market to hold. There are 300,000 investors. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, $b = 0$, and $w = 0.5$.

algorithm is given by

$$\begin{aligned} Q_t^{MB}(a, s_{t\_}) &= \mathbb{E}_t^p\left[\log\left((1-a-w_{t\_})R_f + (a+w_{t\_})R_{m,t+1}\right)\right] \\ &+ \frac{\gamma}{1-\gamma}\mathbb{E}_t^p\left[\log\left((1-a^*-w_{t\_})R_f + (a^*+w_{t\_})R_{m,t+1}\right)\right], \end{aligned} \tag{52}$$

where

$$a^* = \arg\max_a \mathbb{E}_t^p\left[\log\left((1-a-w_{t\_})R_f + (a+w_{t\_})R_{m,t+1}\right)\right]. \tag{53}$$

The updating rule for the probability distribution is the same as that in the main text.

We note three things. First, as before, if $a + w_{t\_}$ is above 100%, it is capped at 100%; and if it is below 0%, it is floored at 0%. Second, the algorithm in (52) assumes that investors are myopic: they think $a^* + w_{t\_}$ is the optimal allocation for all future periods; they do not think about transitions from one state to another. Finally, for each time period $t$, we need update only the model-based $Q$ values for state $s_{t\_}$; the $Q$ values for other states do not matter when investors make decisions at time $t$.

The specifications of model-free learning and model-based learning also allow us to examine the hybrid model, one that assigns equal weight to the two learning systems.

We now implement the above structure and examine how the allocations recommended by the model-free, model-based, and hybrid systems depend on the past 30 years of stock market returns. There are 300,000 investors. At time 0, we randomly select $w_{0\_}$ from the 11 possible allocations $\{0\%, 10\%, \ldots, 100\%\}$. The parameter values are the same as those in

Figure 1 of the main text.

Figure A12 presents the results. While they differ quantitatively from those in Figure 1, its counterpart in the main text, they are qualitatively similar.

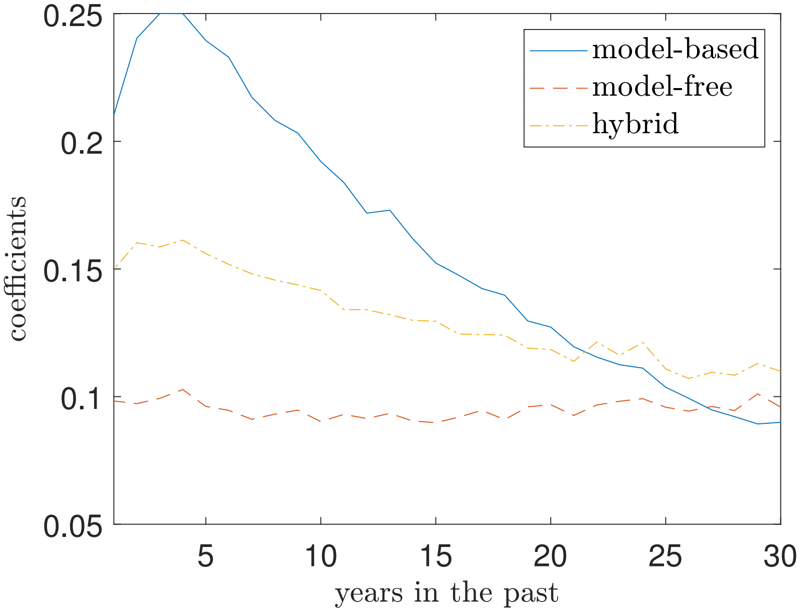

**Figure A12.** Analogous to Figure 1, the graph shows how the allocations recommended by the model-free, model-based, and hybrid systems depend on the past 30 years of stock market returns. In contrast to Figure 1, investors' actions are defined relative to the previous period's allocations. There are 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 $b = 0$. In the case of the hybrid system, $w = 0.5$.

Given that Figure A12 is structurally similar to Figure 1 in the main text, we would expect to observe a frequency disconnect, insensitivity of allocations to beliefs, and experience effects even for this alternative action space. We confirm that this is the case; while there are some quantitative differences, the results are qualitatively the same as in the main text.

## J. Comparison with Models of Inattention

One of the properties of model-free learning is that it generates inertia in investor allocations. It is therefore natural to compare our framework to another framework that is often used to think about inertia in allocations, one based on investor inattention.

We consider three models of inattention. All three take the model-based component of our framework, discard the model-free component, and instead introduce a form of inattention.

In the first approach, each investor updates his beliefs about stock market returns at each date as in equations (17)-(18) in the main text. With probability $p$, he is attentive and also makes an active adjustment to his portfolio allocation: he computes the model-based $Q$ values in (19)-(20) in the main text and then chooses an action probabilistically according to

$$p(a_t = a) = \frac{\exp(\beta Q_t^{MB}(a))}{\sum_{a'} \exp(\beta_t^{MB}(a'))}.$$

However, with probability $1 - p$, he is not attentive, and his allocation drifts passively, so that

$$a_t = \frac{a_{t-1} R_{m,t}}{a_{t-1} R_{m,t} + (1 - a_{t-1})}.$$

In our second approach to modeling inattention, the investor again updates his beliefs in each period according to equations (17)-(18) in the main text. Moreover, in each period, he updates the model-based $Q$ values of all allocations, as in (19)-(20) in the main text. Finally, in each period, he checks whether the expected $Q$ value of his new allocation, if he did make an active choice, exceeds the $Q$ value of his previously-chosen allocation by more than some transaction cost $c$:

$$\frac{\sum_{a'} \exp(\beta Q_t^{MB}(a')) Q_t^{MB}(a')}{\sum_{a'} \exp(\beta Q_t^{MB}(a'))} - Q_t^{MB}(\hat{a}_{t-1}) > c. \tag{54}$$

If this condition is satisfied, the investor chooses an action probabilistically, according to current $Q$ values. Otherwise, his allocation continues to drift passively, so that

$$a_t = \frac{a_{t-1} R_{m,t}}{a_{t-1} R_{m,t} + (1 - a_{t-1})}.$$

In equation (54), $\hat{a}_{t-1}$ is the allocation in the set $\{0\%, 10\%, \ldots, 100\%\}$ that is closest to $a_{t-1}$; since we have $Q$ values only for the 11 feasible allocations, we approximate $Q^{MB}(a_{t-1})$ by $Q^{MB}(\hat{a}_{t-1})$.

Both of the above inattention models assume that the investor can effortlessly update his beliefs in each period. In reality, however, the investor may find it just as effortful to update his beliefs as to change his allocation. We therefore consider a third model, a variant of the first, in which, at each time, the investor is inattentive with probability $1 - p$ and updates neither his beliefs nor his allocation; and with probability $p$, he updates his beliefs and model-based $Q$ values based on all the returns realized since his last belief update and then chooses an action probabilistically based on the $Q$ values.

We now analyze all three models in detail. In particular, we look at their predictions for the main applications in Section 4: the frequency disconnect, the insensitivity of allocations to beliefs, experience effects, and inertia. Throughout, there are 300,000 investors in six cohorts of 50,000 each. The parameter values are $\alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, and $\sigma = 0.2$.

For the first model of inattention, we have studied six possible values of $p$, namely 1, 0.75, 0.5, 0.25, 0.1, and 0.02; the results we present here are for the case of $p = 0.1$. Figure A13, the analog of Figure 4 in the main text, presents the results for the frequency disconnect. Figure A14, the analog of Figure 6, presents results for experience effects. For the parameter values we use here, the sensitivity of allocations to beliefs is 0.703.

For the second inattention model, we have studied seven possible values of c: $-\infty$, 0, 0.01, 0.05, 0.1, 0.15, and 0.2; we illustrate the results here for the case of $c = 0.15$.

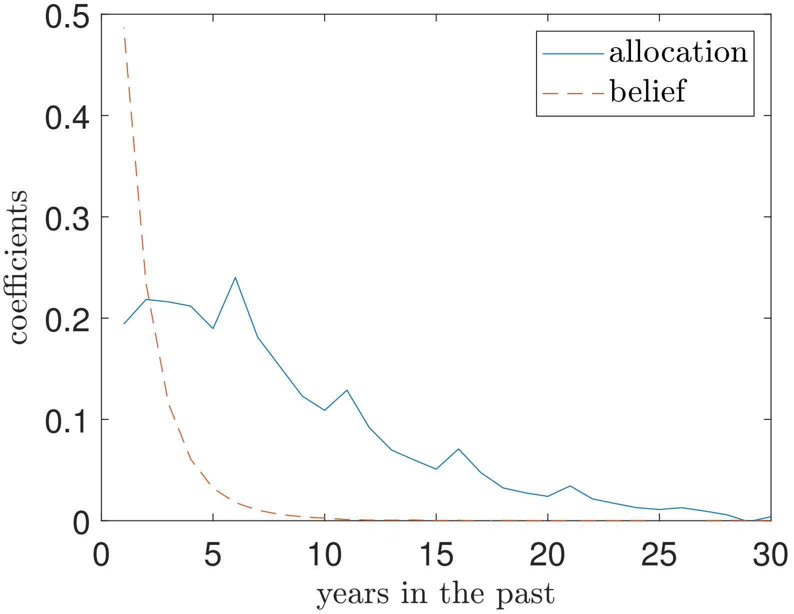

**Figure A13.** Analogous to Figure 4, the graph shows how investors' allocations and beliefs at time 30 depend on the past 30 years of stock market returns. In contrast to Figure 4, investors do not use model-free learning, but rather are inattentive model-based investors. There are 300,000 investors in six cohorts of 50,000 each. The parameter values are $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $p = 0.1$.

Figure A15, the analog of Figure 4, presents the results for the frequency disconnect; Figure A16, the analog of Figure 6, presents the results for the case of experience effects. For $c = 0.15$, the sensitivity of allocations to beliefs is 1.062.

The results for these two inattention models are similar: they can generate a frequency disconnect and insensitivity of allocations to beliefs. Interestingly, though, they make a prediction about experience effects that is quite different from that of our framework, namely that, if an investor enters financial markets at time $t$, his allocation at time $T$ will typically put *more* weight on the most recent return he did not experience, $R_{m,t}$, than on the first return he did experience, $R_{m,t+1}$. The reason is that, when an investor enters financial markets, he is paying attention, and so takes account of the return just before he enters, $R_{m,t}$. However, one year later, he may not be paying attention and may therefore not account for the return at that time, $R_{m,t+1}$. By contrast, our Section 2 framework makes the opposite prediction, one that is more in line with the evidence on experience effects, namely that the investor will put more weight on $R_{m,t+1}$ than on $R_{m,t}$.

Finally, we analyze the third inattention model, one in which the investor is inattentive in updating both beliefs and allocations. We find that the results here are similar to those of the first two inattention models on most dimensions: this model also has trouble generating realistic experience effects. However, it differs from our framework in its predictions in an additional important way: it is less able to generate insensitivity of allocations to beliefs; for $p = 0.1$, it generates a sensitivity of 1.769, which is more than double the insensitivity for the first inattention model, namely 0.703.

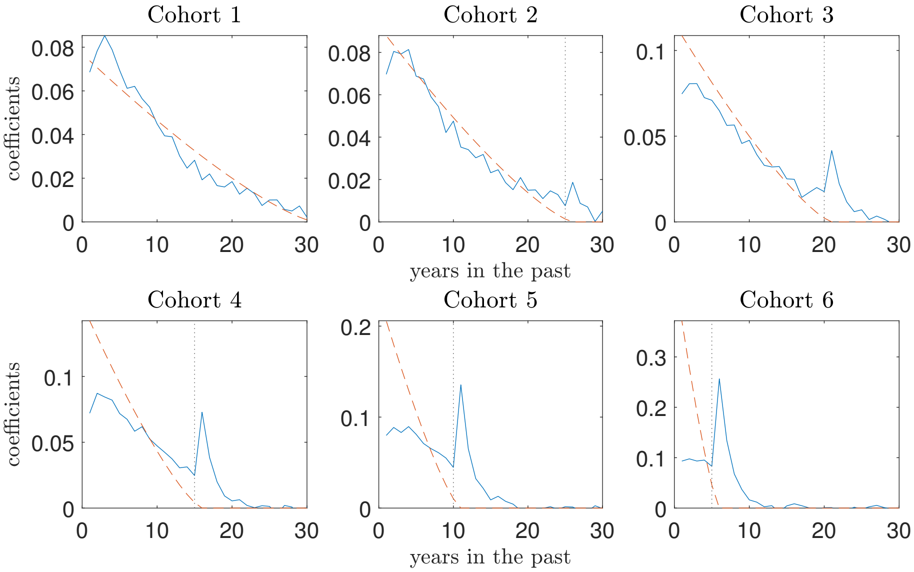

**Figure A14.** Analogous to Figure 6, the graph shows how the allocations of each of the six cohorts depend on past stock market returns. In contrast to Figure 6, investors do not use model-free learning, but rather are inattentive model-based investors. There are 300,000 investors in six cohorts of 50,000 each. The parameter values are $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $p = 0.1$.

## K. Parameter Estimation

In this section, we describe the procedure that we use 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 $\beta$; 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 $\bar{\alpha}^{MF}$, $\beta$, 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 $b = 0.0577$.[5]

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

[5] 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.

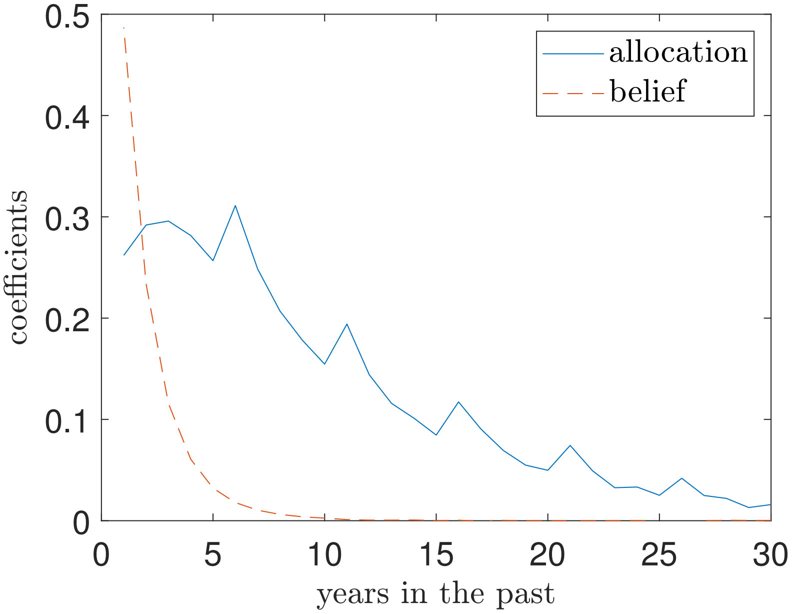

**Figure A15.** Analogous to Figure 4, the graph shows how investors' allocations and beliefs at time 30 depend on the past 30 years of stock market returns. In contrast to Figure 4, investors do not use model-free learning, but rather are inattentive model-based investors. There are 300,000 investors in six cohorts of 50,000 each. The parameter values are $L = T = 30$, $\alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $c = 0.15$.

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, $\alpha_{+}^{MB}$ and $\alpha_{-}^{MB}$, 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

$$(c_1 - 0.127)^2 + (\frac{c_2}{c_1} - 0.29)^2 + (\frac{c_3}{c_2} - 0.77)^2. \tag{55}$$

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

With this value of $\bar{\alpha}^{MB}$ in hand, we search for values of $\bar{\alpha}^{MF}$, $\beta$, 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 $\bar{\alpha}^{MF}$, $\beta$, and $w$, we can compute this coefficient, $d$, in simulated data from our framework.

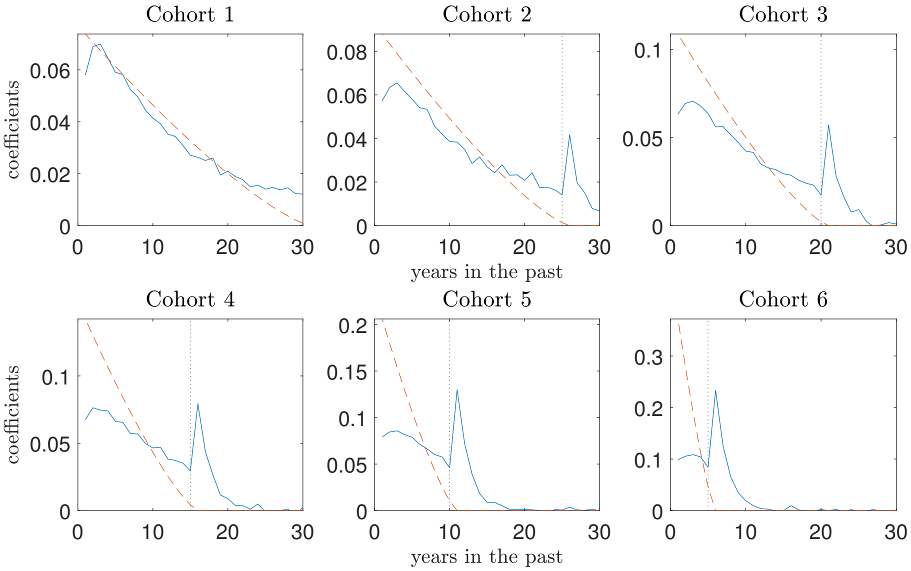

**Figure A16.** Analogous to Figure 6, the graph shows how the allocations of each of the six cohorts depend on past stock market returns. In contrast to Figure 6, investors do not use model-free learning, but rather are inattentive model-based investors. There are 300,000 investors in six cohorts of 50,000 each. The parameter values are $L = T = 30$, $\alpha^{MB}_{\pm} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $c = 0.15$.

Due to computational constraints, these simulated data are now based on 60,000 investors in six cohorts of 10,000 each.

Our second target is the functional form in (25) 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 $\bar{\alpha}^{MF}$, $\beta$, 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 $\text{SSE}_1$, the sum of squared errors for cohort 1. In a similar way, we compute $\text{SSE}_i$ 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

$$(d-1)^2 + \sum_{i=1}^{6} \text{SSE}_i \tag{56}$$

and identify the parameter values that minimize this quantity. The first term in (56) targets the empirical sensitivity of allocations to beliefs, while the second term targets the empirical experience effects. We find that the parameter values that minimize (56) are $\bar{\alpha}^{MF} = 0.26$, $\beta = 20$, and $w = 0.38$. The estimate of $w$ in particular indicates that our framework can best match the empirical facts when it puts substantial weight on both the model-free and model-based systems.

### L. System Performance and an Analysis of State Dependence

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. To illustrate this, we replace the i.i.d return structure of Section 2 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. An analysis of this case necessarily introduces a state variable, namely the asset's past return.

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

$$S_{m,t} = (1-\theta)\sum_{i=0}^{t-1}\theta^i R_{m,t-i} + \theta^t S_{m,0}, \tag{57}$$

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

$$\log R_{m,t+1} = \mu_L + \sigma\varepsilon_{t+1}, \tag{58}$$

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

$$\log R_{m,t+1} = \mu_H + \sigma\varepsilon_{t+1}, \tag{59}$$

where $\mu_H$ 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

$$\log R_{m,t+1} = \mu_M + \sigma\varepsilon_{t+1}, \tag{60}$$

where $\mu_M$ takes a moderate value; we call this the Moderate state, $M$. In each case, $\varepsilon_{t+1}$ 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* able to recognize and observe the three states, his learning algorithms are different. For model-free learning, the $Q$ values are updated according to

$$Q_{t+1}^{MF}(s_t, a) = Q_t^{MF}(s_t, a) + \alpha_{t,\pm}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(s_{t+1}, a') - Q_t^{MF}(s_t, a)] \tag{61}$$

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

$$p_{t+1}(R_m = R, s_t) = p_t(R_m = R, s_t) + \alpha_{t,\pm}^{MB}[1 - p_t(R_m = R, s_t)] \tag{62}$$

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:

$$\begin{aligned}
Q_t^{MB}(s_t = L, a) &= \mathbb{E}_t^{p,L} \log((1-a)R_f + aR_{m,t+1}) + \gamma(\chi^{LH}V^H + \chi^{LM}V^M + \chi^{LL}V^L), \\
Q_t^{MB}(s_t = M, a) &= \mathbb{E}_t^{p,M} \log((1-a)R_f + aR_{m,t+1}) + \gamma(\chi^{MH}V^H + \chi^{MM}V^M + \chi^{ML}V^L), \\
Q_t^{MB}(s_t = H, a) &= \mathbb{E}_t^{p,H} \log((1-a)R_f + aR_{m,t+1}) + \gamma(\chi^{HH}V^H + \chi^{HM}V^M + \chi^{HL}V^L), \qquad (63)
\end{aligned}$$

where $\mathbb{E}_t^{p,s}$ represents the investor's perceived return distribution in state $s$ at time $t$, $\chi^{s_1,s_2}$ 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

$$Q_t^{HYB}(s_t, a) = (1-w)Q_t^{MF}(s_t, a) + wQ_t^{MB}(s_t, a). \tag{64}$$

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

$$p(s_t, a_t = a) = \frac{\exp[\beta Q_t^{HYB}(s_t, a)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(s_t, a')]}. \tag{65}$$

Equation (65) shows that the values of $\chi^{s_1,s_2}$ 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 (65) that contains $\chi^{s_1,s_2}$ and $V^s$ is cancelled out by the same term in the denominator.

We now present some numerical analysis. The parameters $\sigma$, $\alpha_{t,\pm}^{MF}$, $\alpha_{t,\pm}^{MB}$, $\gamma$, $w$, and $\beta$ 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.

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. 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.

## M. 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

$$p(a_t = a) = \frac{\exp[\beta Q_{t-1}^{MF}(a)]}{\sum_{a'} \exp[\beta Q_{t-1}^{MF}(a')]}, \tag{66}$$

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

$$Q_t^{MF}(a_{t-1}) = Q_{t-1}^{MF}(a_{t-1}) + \alpha_{t-1,\pm}^{MF} \left[\log R_{p,t} + \gamma Q_{t-1}^{MF}(a_t) - Q_{t-1}^{MF}(a_{t-1})\right]. \tag{67}$$

Analogous to the analysis in Section 3, 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 A17 presents the results. We make 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.

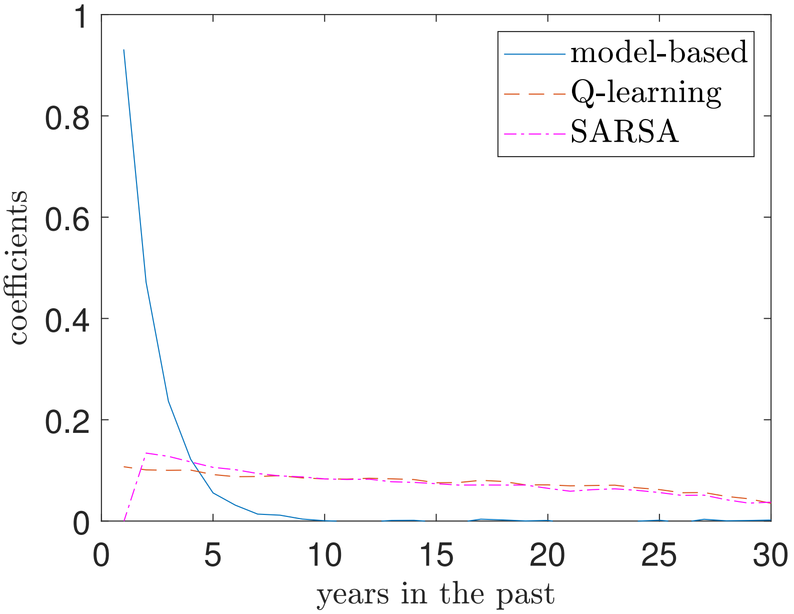

**Figure A17.** We run a regression of investors' allocations to the stock market $a_T$ at time $T$ on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients for three cases: model-based learning; model-free Q-learning; and model-free 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.