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

Nicholas Barberis and Lawrence Jin

October 2021*

**Abstract**

In the past decade, researchers in psychology and neuroscience studying human decision-making have increasingly adopted a framework that combines two systems, namely "model-free" and "model-based" learning. We import this framework into a simple financial setting, study its properties, and link it to a wide range of applications. We show that it provides a foundation for extrapolative demand and experience effects; resolves a puzzling disconnect between investor allocations and beliefs in both the frequency domain and the cross-section; helps explain the dispersion in stock market allocations across investors as well as the inertia in these allocations over time; and sheds light on the persistence of household investment mistakes. More broadly, the framework offers a way of thinking about individual behavior that is grounded in recent evidence on the computations that the brain undertakes when estimating the value of a course of action.

*This is a preliminary draft – comments are welcome. The authors' affiliations are Yale University and Caltech, respectively; their e-mails are nick.barberis@yale.edu and lawrence.jin@caltech.edu. We are grateful to Cary Frydman, Ulrike Malmendier, Antonio Rangel, Mike Woodford, and seminar participants at Arizona State University, Caltech, the University of California Los Angeles, the University of Pennsylvania, Yale University, and the Behavioral Economics Annual Meeting (BEAM) for very useful feedback. We are also grateful to Colin Camerer, Nathaniel Daw, Sam Gershman, John O'Doherty, and members of their lab groups for very helpful discussions about the psychological concepts in the paper. Steven Ma provided excellent research assistance.

# 1 Introduction

In the past decade, psychologists and neuroscientists have increasingly embraced a new framework for thinking about human decision-making in experimental settings – a framework that differs from that used by economists. The 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 an economic setting – a simple portfolio-choice problem where investors allocate between a risk-free asset and a risky asset – study its properties, and show that it is helpful for thinking about a range of facts in finance.[1]

The model-free and model-based learning algorithms operate in the following dynamic setting. At each time, after observing the state of the world, an individual takes an action. In the next period, as a consequence, he receives a reward and arrives in a possibly new state of the world. His goal is to choose an action at each time to maximize the long-term sum of rewards.

The model-free and model-based systems both try to solve this problem by estimating a quantity denoted by $Q(s, a)$, the value of taking action $a$ in state $s$. However, they do so in different ways. The model-free system is especially different from the framework used by economists in that, as its name indicates, it does not use a model of the world; in other words, it does not use any information about the probabilities of future states and rewards. Instead, it learns from experience. After taking the action $a$ in state $s$ and observing the subsequent reward, it updates its estimate of $Q(s, a)$ by way of two important quantities: a reward prediction error, which, loosely speaking, is the difference between the reward the individual received and the reward he expected; and a learning rate, which controls the extent of the updating. This model-free framework has been increasingly adopted by psychologists and neuroscientists because of evidence that it reflects actual computations performed by the brain: numerous studies have found that neurons in the brain encode the reward prediction error used by model-free learning.[2]

The model-based system is much closer to the frameworks used by economists in that, as in almost all economic models, it makes use of a probability distribution of future rewards and states conditional on past actions and states. There are many possible model-based approaches; we use one that is often adopted in research in psychology and that, like the model-free system, has neuroscientific support. In this framework, after taking an action

[1] An early paper on this framework is Daw, Niv, and Dayan (2005). Useful reviews include Balleine, Daw, and O'Doherty (2009) and Daw (2014).

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

and observing the subsequent reward and state, the individual increases the probability he assigns to that reward and state while downweighting the probabilities of other outcomes. To do the updating, he again uses a learning rate and a prediction error, often called a state prediction error, which measures how surprising the realized state and reward are. As with the reward prediction error, there is evidence that the brain computes such state prediction errors (Glascher et al., 2010).

Recent research in psychology argues that, to make decisions, people use these two systems in combination: they take a weighted average of the $Q(s, a)$ values produced by each of the model-free and model-based systems and use the resulting "hybrid" $Q(s, a)$ to make a choice (Glascher et al., 2010; Daw et al., 2011).

In this paper, we import this framework into an economic setting, study its properties, and use it to account for a range of facts about investor behavior. The setting we consider is a simple portfolio-choice problem where an individual allocates money between a risk-free asset and a risky asset in order to maximize the expected log utility of wealth at some future horizon. This problem fits the canonical context in which model-free and model-based learning algorithms are applied.

Another difference between the model-free and model-based systems – one that is reflected in our framework – is that the model-free system is likely to operate over a more limited time range: because it learns from experienced rewards, it is in operation only when the individual is actively interacting with the environment – for example, only when he is actively experiencing financial markets. By contrast, the model-based framework is trying to build a model of how rewards depend on states and actions, and it can do so using data from before the individual started experiencing the environment – for example, from before he started actively investing.

We begin by analyzing the properties of our framework. We focus on the model-free system – the more novel part of our framework – and on its interaction with the model-based system. We find that, while the model-free algorithm is as simple if not simpler than its model-based counterpart, it leads to richer predictions as well as novel economic intuitions.

We start by looking at how the stock market allocation proposed by each of the model-free and model-based systems depends on past stock market returns. The model-based allocation puts weights on past returns that are positive and that decline for more distant past returns. For many parameter values, the model-free system also recommends an allocation that puts positive and declining weights on past returns. However, relative to the model-based system, it puts substantially more weight on distant past returns. This is because it learns slowly: at each time, it learns primarily about the action the individual is currently taking. For some parameterizations, it can even put more weight on distant returns than on recent returns. Moreover, the relative weight it assigns to recent as opposed to distant returns is affected by factors that play no role in the model-based allocation – factors such as the discount rate and the number of allocation choices available to the investor. We also find that the model-free system generates more inertia in investor allocations over time, and that while each of the two systems can lead to both underreaction and overreaction, the model-free system exhibits substantially more underreaction.

We then use our framework to shed light on a range of facts about investor behavior. A prominent idea, motivated by empirical evidence, is that investors have "extrapolative" demand: their demand for a risky asset is a weighted average of past returns, with more weight on more recent returns. Our framework offers a new foundation for extrapolative demand, one rooted in the mechanics of the model-free system. It further posits that this demand is the sum of two distinct components operating at different frequencies – a model-based component which puts high weight on recent returns and a model-free component that puts substantial weight even on distant past returns.

Our framework also provides a foundation for experience effects – specifically, for the empirical finding of Malmendier and Nagel (2011) that an individual's allocation to the stock market can be explained in part by a weighted average of the market returns he has personally experienced, with substantial weight on even distant past experienced returns but less weight on returns he has not experienced. Our framework can capture this by way of its model-free component. As noted above, the model-free system puts substantial weight on distant past experienced rewards. Moreover, since the model-free system is in operation only when the individual is actively investing, it puts no weight on returns he has not directly experienced.

Our framework can also resolve a puzzling disconnect between investors' stock market allocations and investors' beliefs. Greenwood and Shleifer (2014), among others, use survey data to show that investor beliefs about future stock market returns depend primarily on recent past market returns. However, Malmendier and Nagel (2011) find that investors' allocations to the stock market depend significantly even on distant past market returns. Two aspects of our framework allow us to resolve this. First, only one of the two systems, the model-based system, has an explicit role for beliefs. Second, the other system, the model-free system, proposes allocations that depend even on distant past returns. As a result, when an individual is surveyed about his beliefs regarding future returns, he consults the model-based system and gives an answer that depends primarily on recent past returns. However, when he chooses an allocation, he uses both the model-free and model-based systems and hence chooses an action that depends significantly even on distant past returns. Through a similar mechanism, our framework can also explain the low sensitivity of allocations to beliefs in the cross-section of investors documented by Giglio et al. (2021).

We show that the framework can also help to address other empirical facts, including the large cross-sectional dispersion in investor allocations to the stock market; the individual-level inertia in these allocations over time; and the widespread non-participation in the stock market among U.S. households.

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

Beyond the applications discussed above, our framework also points to some broader themes. First, by way of a simple estimation exercise, we find that the parameter values that best fit the data put substantial weight on the model-free system. It is striking that investors would put a lot of weight on a system that uses no information about the statistical structure of returns, even though such information is in principle available. This may be an indication that many people have a poor sense of the structure of asset returns, and therefore fall back on a system – the model-free system – that requires no knowledge of this structure. It may also be a sign of how fundamental the model-free system is to human decision-making.

Second, when economists try to explain investors' differing stock market allocations, they typically look to differences in risk aversion or to differences in beliefs about future returns or risk. Our framework suggests that a deeper driver of differing allocations may be investor learning rates. As such, a potentially useful direction for future research is to measure these learning rates and to connect them to investor behavior.

Third, many papers in economics start with a dynamic investment problem, use mathematical or numerical techniques to derive the value function, and then use this value function to interpret observed behavior. However, this line of research rarely explains how an individual might actually come to act in the way described by the value function – a natural question to ask, given that few people know how to solve the Bellman equations that characterize the value function. By contrast, in this paper, we try to explain individual behavior with a framework that is rooted in, and consistent with, algorithms that the brain appears to use when estimating the value of different courses of action.

Model-free learning algorithms are of interest not only to psychologists and neuroscientists, but also to computer scientists, albeit for a different purpose. Computer scientists see these algorithms as a powerful tool for solving difficult dynamic problems (Sutton and Barto, 2019). For example, these algorithms have been embedded in computer programs that have achieved world-beating performance in complex games such as Backgammon and Go. Psychologists and neuroscientists, by contrast, are interested in these algorithms because they see them as good models of how animals and humans actually behave. In this paper, we take the psychologists' perspective: we are proposing that these algorithms can shed light on the behavior of real-world investors.

The full name of model-free learning is model-free reinforcement learning. Reinforcement learning is heavily used in both psychology and neuroscience – and, as described above, in some areas of computer science. However, it has a much smaller footprint in economics and finance, where model-based frameworks dominate instead. Nonetheless, our approach does have antecedents in economics – most notably in research in behavioral game theory on how people learn what actions to take in strategic settings (Camerer, 2003, Ch. 6). For example, one important idea in this line of research, namely Camerer and Ho's (1999) experience-weighted attraction learning, combines reinforcement and model-based learning in a way that is reminiscent of the hybrid model we consider below.

Our analysis differs from this earlier work, most notably in the models we use and in the setting we consider. It is not surprising that our models are different – the specific framework we adopt was developed in large part after the key papers in behavioral game theory were written. Moreover, the financial setting we focus on is very different from the experimental and strategic settings considered in prior work, and leads to entirely new applications.

In Section 2, we describe the model-free and model-based systems and show how they can be applied to a simple portfolio-choice problem. In Section 3, we analyze the properties of the framework and present an example that illustrates the mechanics of the two systems. In Section 4, we connect the framework to a range of applications. Section 5 discusses some broader themes, while Section 6 concludes.

## 2 Models

Researchers in cognitive psychology and decision neuroscience are increasingly adopting a new framework for studying human behavior in experimental settings, one that differs from that used by economists. The framework has two components, known as "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 incorporating it into a financial setting.

Model-free and model-based learning algorithms are intended to solve decision problems of the following form. Time is discrete and indexed by $t = 0, 1, 2, 3, \ldots$ At time $t$, the state of the world is denoted by $s_t$ and the individual takes an action $a_t$. As a consequence of taking the action $a_t$ in state $s_t$ at time $t$, the individual receives a reward $R_{t+1}$ at time $t+1$ and arrives in state $s_{t+1}$ at that time. The joint probability of $s_{t+1}$ and $R_{t+1}$ conditional on $s_t$ and $a_t$ is denoted $p(s_{t+1}, R_{t+1}|s_t, a_t)$. The environment has a Markov structure: the probability of $(s_{t+1}, R_{t+1})$ depends only on $s_t$ and $a_t$.

In a finite-horizon setting, the individual's goal is to maximize the expected sum of rewards:

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

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

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

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

Economists almost always tackle a problem of this type using dynamic programming. Under this approach, we solve for the time $t$ value function $V(s_t)$ – the expected sum of discounted future rewards, under the optimal policy, conditional on being in state $s_t$ at time $t$. To do this, we write down the Bellman equation that $V(s_t)$ satisfies, and, with the probability distribution $p(s_{t+1}, R_{t+1}|a_t, s_t)$ in hand, we solve the equation, either analytically or numerically. The solution is sometimes used for "normative" purposes – to tell the individual how he *should* act – and sometimes for "positive" purposes, to explain observed behavior.[3]

Whether we have in mind a positive or a normative application, the dynamic programming approach has its limits. For many normative applications, it may be difficult to compute

[3]The dynamic programming methodology is used by both rational and behavioral approaches. The only difference is that, under a behavioral approach, the individual may have an incorrect perception of the probability distribution $p(\cdot)$.

the probability distribution $p(\cdot)$; and even if we have a good sense of this distribution, it may be hard to solve the Bellman equation for the value function $V(s_t)$.[4]

For "positive" applications, where we are trying to describe how people actually behave, the dynamic programming approach raises yet more questions. Again, it is often difficult, even for professional economists, to determine the distribution $p(\cdot)$ and to then solve the Bellman equation for the value function. As such, it is not clear how ordinary individuals would be able to do so. Economists have long suggested that people act "as if" they have solved the Bellman equation, even if they have not done so explicitly: just as people can play billiards without knowing the physics of ball movement, so they may be able to intuit their way to the optimal solution of an economic decision problem.

The "as if" argument faces at least two difficulties. First, in many economic settings, people's behavior does not appear to be fully rational (Barberis and Thaler, 2003). As such, we cannot presume that people are able to intuit their way to the optimal solutions of economic problems. Second, under the "as if" approach, people's behavior remains a black box, in that we do not know how they have come to decide on a particular course of action. It seems preferable to try to understand individual behavior using a framework that is rooted in, and consistent with, the actual computations the brain performs when making a decision. In this paper, we take a step in this direction.

The difficulties in using dynamic programming even for some normative applications has led computer scientists to develop alternative algorithms for solving the problems in (1) and (2) (Sutton and Barto, 2019). An important subset of these is model-free reinforcement learning algorithms. As their name suggests, these are algorithms that tackle the problems in (1) and (2) without a "model" of the world, in other words, without using any information about the probability distribution $p(\cdot)$. The important finding that motivates this paper is that some of these algorithms are not only useful ways of solving the problems in (1)-(2), but also appear to reflect actual computations performed by the brain.[5] The model-free algorithms most commonly used by psychologists to understand decision-making in experimental settings are Q-learning and SARSA. In this paper, we use Q-learning. We have repeated the main parts of our analysis for SARSA and obtain similar results.[6]

[4]For example, while the problem of determining optimal play in the card game Blackjack fits into the setting laid out above, it is very hard to solve this problem using dynamic programming, both because of the difficulty of computing the distribution $p(\cdot)$ and because of the large number of states (Sutton and Barto, 2019, Ch. 5).

[5]See Montague, Dayan, and Sejnowksi (1996), Schultz, Dayan, and Montague (1997), McClure, Berns, and Montague (2003), and O'Doherty et al. (2003).

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

In Section 2.1, we present the model-free Q-learning algorithm; in Section 2.3, we lay out a model-based learning algorithm; and in Section 2.4, we show how the model-free and model-based algorithms are combined. In Section 2.2, we describe a simple portfolio-choice setting that fits the structure of the decision problem above. For much of the paper, we will explore the properties and possible applications of model-free and model-based learning in this financial setting.

## 2.1 Model-free learning

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

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

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

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 results in a reward $R_{t+1}$ at time $t+1$ and brings us to state $s_{t+1}$ at that time. Suppose also that, at time $t$, the algorithm's initial estimate of $Q^*(s,a)$ is $Q^{\text{old}}(s,a)$. At time $t+1$, after observing the reward $R_{t+1}$, its estimate of $Q^*(s,a)$ is updated as follows:

$$Q^{\text{new}}(s,a) = Q^{\text{old}}(s,a) + \alpha_t^{MF}[R_{t+1} + \gamma \max_{a'} Q^{\text{old}}(s_{t+1},a') - Q^{\text{old}}(s,a)], \tag{4}$$

where $\alpha_t^{MF}$ is known as the learning rate – the superscript stands for model-free – and the term in square brackets is an important quantity known as the reward prediction error (RPE): the realized value of taking the action $a$ relative to its previously anticipated value.

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 $Q^{\text{old}}(s,a_t)$. Rather, it chooses an action probabilistically, where the probability of choosing a given action is an increasing function of its $Q$ value:

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

This probabilistic choice, often known as a "softmax" approach, serves an important purpose: it encourages the algorithm to "explore," in other words, to try an action other than the one that currently has the highest $Q$ value in order to see whether this other action has an even higher $Q$ value. In the limit as $\beta \to \infty$, the algorithm chooses the action with the highest $Q$ value; in the limit as $\beta \to 0$, it chooses an action randomly. The parameter $\beta$ 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.

The algorithm is initialized at time 0 by setting $Q(s,a) = 0$ for all $s$ and $a$. Consistent with (5), the time 0 action is chosen randomly from the set of possible actions. The process then proceeds according to equations (4) and (5).

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

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

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

$$Q^{\text{new}}(s,a) = (1 - \alpha_t^{MF}) Q^{\text{old}}(s,a) + \alpha_t^{MF}[R_{t+1} + \gamma \max_{a'} Q^{\text{old}}(s_{t+1}, a')], \tag{7}$$

we see that the Q-learning algorithm is taking an estimate of the right-hand side of (6) and then updating $Q^{\text{old}}(s,a)$ in the direction of this estimate to an extent determined by the learning rate $\alpha_t^{MF}$. Specifically, it proxies for the expected reward $E_t(R_{t+1})$ in (6) by the realized reward $R_{t+1}$ and for $E_t[\max_{a'} Q^*(s_{t+1}, a')]$ by $\max_{a'} Q^{\text{old}}(s_{t+1}, a')$.

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

$$\begin{aligned} Q^{\text{new}}(s,a) &= Q^{\text{old}}(s,a) + \alpha^{MF}_{t,+}(\text{RPE}) \qquad \text{for RPE} \geq 0 \\ Q^{\text{new}}(s,a) &= Q^{\text{old}}(s,a) + \alpha^{MF}_{t,-}(\text{RPE}) \qquad \text{for RPE} < 0. \end{aligned} \tag{8}$$

In what follows, we also adopt this modification.

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

## 2.2 A portfolio-choice setting

In Section 2.3, we lay out a model-based algorithm to complement the model-free algorithm described above. Before we do so, we first describe the financial problem that we will apply both algorithms to – a problem that fits the setting specified above.

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

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

At each time $t$, an investor chooses the fraction of his wealth that he allocates to the risky asset; this corresponds to the "action" in the framework of Section 2.1, so we use the

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

notation $a_t$ for it.[8] The investor's goal is to maximize the expected log utility of wealth at some future horizon determined by his liquidity needs. Because the timing of these liquidity needs is uncertain, he does not know in advance how far away this horizon is. Specifically, at time 0, the investor enters financial markets. If, coming into time $t \geq 1$, he is still present in financial markets, then, with probability $1-\gamma$, where $\gamma \in [0,1)$, a liquidity shock arrives at time $t$. In that case, he exits financial markets and receives log utility from his wealth at time $t$. A simple calculation – see the Appendix – shows that the investor's implied objective is to solve

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

where $R_{p,t}$, the investor's gross portfolio return between time $t-1$ and $t$, is given by

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

Comparing (2) and (10), we see that this portfolio problem maps directly into the framework of Section 2.1: the generic reward $R_t$ in equation (2) now has a concrete form, namely the log portfolio return, $\log R_{p,t}$.

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

$$a^* = \arg\max_a E_t \log((1-a)R_f + aR_{m,t+1}). \tag{11}$$

The fact that the problem in (10) can be solved mathematically does not necessarily mean that real-world investors will be able to find their way to the solution in (11). 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 (11) will *not* be intuitive, in that it involves reducing exposure to the stock market after the market has performed well and increasing exposure to the stock market after the market has performed poorly – actions that will feel unnatural to many investors.

If investors are unable to explicitly compute the solution to the problem in (10), they may instead rely on a model-free algorithm like Q-learning, for at least two reasons. First, the model-free algorithm does not use any information about the structure of asset returns; this makes it appealing to investors who feel uninformed about this structure. Second, the model-

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

free algorithm is thought to be a fundamental tool for human decision-making in general; as such, it is likely to exert at least some influence in any setting a person encounters.

How can Q-learning be applied to the above problem? In principle, we could apply equation (7) directly. However, it is natural to start with a simpler case – the case with no state dependence, so that $Q(s,a)$ is replaced by $Q(a)$. Even this simple case has rich implications that shed light on empirical facts, and so it will be our main focus. The psychological interpretation of removing the state dependence is that it is a simplification on the part of investors. Indeed, neuroscientists have argued that, in an effort to speed up learning, the brain does try to simplify the state structure when implementing its learning algorithms (Collins, 2018).[9]

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

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

– if the investor chooses the allocation $a$ at time $t$ and then continues optimally from the next period on. Suppose that, at time $t$, the individual chooses the allocation $a$ and observes the reward – the log portfolio return, $\log R_{p,t+1}$ – at time $t+1$. He then updates his model-free estimate of $Q^*(a)$ from $Q_t^{MF}(a)$ to $Q_{t+1}^{MF}(a)$ according to

$$Q_{t+1}^{MF}(a) = Q_t^{MF}(a) + \alpha_{t,\pm}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)], \tag{12}$$

where $\alpha_{t,\pm}^{MF}$ equals $\alpha_{t,+}^{MF}$ if the reward prediction error is positive and $\alpha_{t,-}^{MF}$ otherwise. At any time $t$, he chooses his allocation $a_t$ probabilistically, according to

$$p(a_t = a) = \frac{\exp[\beta Q_t^{MF}(a)]}{\sum_{a'}\exp[\beta Q_t^{MF}(a')]}. \tag{13}$$

The exploration embedded in (13) is central to the model-free algorithm and an integral part of how psychologists think about human behavior. By contrast, the term is rarely used in economics or finance. Nonetheless, many actions in financial settings can be thought of as forms of exploration – for example, any time an individual tries a strategy that is new to him, such as investing in a stock in a different industry or foreign country, or in an entirely

[9] It is tempting to justify the removal of the state dependence by saying that investors understand that asset returns are i.i.d., which means that the allocation problem has the same form at each time $t$, so that dropping the dependence of $Q(s,a)$ on $s$ is justified. However, we cannot use this argument because the model-free system does not know that returns are i.i.d.; by its nature, it does not have a model of the environment.

new asset class. In our setting, with one risk-free and one risky asset, exploration can be thought of as the investor choosing a different allocation to the stock market than before in order to learn more about what it feels like to do so.

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

$$Q^*(a) = E\log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma}E\log((1-a^*)R_f + a^*R_{m,t+1}), \tag{14}$$

where $a^*$ is defined in (11).

In the basic model-free algorithm in (12), after taking action $a_t = a$ at time $t$, only the $Q$ value of action $a$ is updated. It is natural to ask whether the algorithm can generalize from its experience of taking the action $a$ in order to also update the $Q$ values of other actions. A large literature in computer science has studied this kind of model-free generalization (Sutton and Barto, 2019, Chs. 9-13). As important for our purposes, research in psychology suggests that the human model-free system also engages in generalization (Shepard, 1987). We therefore incorporate generalization into our framework.

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

$$Q^{MF}_{t+1}(\widehat{a}) = Q^{MF}_t(\widehat{a}) + \alpha^{MF}_{t,\pm}\kappa(\widehat{a})[\log R_{p,t+1} + \gamma \max_{a'} Q^{MF}_t(a') - Q^{MF}_t(a)], \tag{15}$$

where

$$\kappa(\widehat{a}) = \exp(-\frac{(\widehat{a}-a)^2}{2b^2}). \tag{16}$$

In words, after observing the reward prediction error for action $a$ and updating the $Q$ value of action $a$, 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 $\alpha^{MF}_{t,\pm}\kappa(\widehat{a})$ – and one that is all the lower, the more different $\widehat{a}$ is from $a$, to an extent determined by the Gaussian function in (16).[10]

[10] Our generalization algorithm is consistent with research in psychology which identifies similarity as a

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 (16), normalized to form a probability distribution, has the same standard deviation as a uniform distribution with width 0.2 – specifically, 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 $b$.

The applications we discuss in Section 4 do not rely on generalization: they follow even for the basic model-free algorithm in (12). We incorporate generalization because it is psychologically reasonable and because it has interesting implications of its own, some of which will improve our framework's fit with the data.

We emphasize that the Q-learning algorithm above, with or without generalization, does not use any information about the distribution of asset returns in (9): by its model-free nature, it does not have a model of the environment. In fact, the algorithm has no idea what a "risk-free asset" or the "stock market" are. It is simply choosing an action – some combination of these unfamiliar objects – seeing what reward it delivers, and then updating the values of the chosen action and of actions similar to it.[11]

### 2.3 Model-based learning

Current research in psychology and decision neuroscience uses a framework in which individual decisions are guided by both model-free and model-based learning. Model-based frameworks, as their name indicates, try to build a model of the environment – for example, in our setting, a model of stock market returns. There are various possible model-based frameworks. Which one should we choose? Our goal in this paper is to see if algorithms commonly used by psychologists can explain behavior in economic settings. We therefore take as our model-based framework one that, like the model-free framework of Section 2.1,

critical driver of generalization (Shepard, 1987). It is also used in computer science, where it is known as interpolation-based Q-learning (Szepesvari, 2010, Ch. 3.3.2). Computer scientists typically use more sophisticated forms of generalization such as function approximation with polynomial, Fourier, or Gaussian basis functions (Sutton and Barto, 2019, Ch. 9). We have also implemented this more complex generalization and obtain similar results. For our main analysis, we therefore stick with the simpler form of generalization in (15)-(16).

[11] Our generalization algorithm treats a 70% algorithm as "similar" to an 80% algorithm. Our interpretation of this is not that the model-free algorithm understands that a 70% and an 80% allocation are likely to lead to similar outcomes; again, the algorithm does not use information about the structure of the task. Rather, the interpretation is simply based on numerical "topology": the number 70 is closer to 80 than to 20.

is based on an algorithm used extensively by psychologists and that is supported by neural evidence from decision-making experiments.

In the framework we consider, an investor learns the distribution of stock market returns over time by observing realized market returns. At each date, he updates the probabilities of different returns using prediction errors analogous to the reward prediction errors of Section 2.1 that are sometimes referred to as "state prediction errors." Specifically, suppose that the investor observes a stock market return $R_{m,t+1} = R$ at time $t+1$ and that, before observing the return, the prior probability he assigned to it occurring was $p^{\text{old}}(R_m = R)$. At time $t+1$, he updates the probability of this return as:

$$p^{\text{new}}(R_m = R) = p^{\text{old}}(R_m = R) + \alpha_t^{MB}[1 - p^{\text{old}}(R_m = R)], \tag{17}$$

where $\alpha_t^{MB}$ is the model-based learning rate that applies from time $t$ to time $t+1$. The term $1 - p^{\text{old}}(R_m = R)$ can be thought of as a prediction error: the investor's prior estimate of the probability of the return equaling $R$ was $p^{\text{old}}(R_m = R)$; when the return is realized, the probability of it equaling $R$ is 1. After this update, the individual scales the probabilities of all other returns down by $1-\alpha_t^{MB}$ 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, $p^{\text{old}}(R_m = R) = 0$, which simplifies (17) to

$$p^{\text{new}}(R_m = R) = \alpha_t^{MB}.$$

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

$$\begin{aligned}
&(R_1, 1) \\
&(R_1, 1 - \alpha^{MB}; R_2, \alpha^{MB}) \\
&(R_1, (1 - \alpha^{MB})^2; R_2, \alpha^{MB}(1 - \alpha^{MB}); R_3, \alpha^{MB}) \\
&(R_1, (1 - \alpha^{MB})^3; R_2, \alpha^{MB}(1 - \alpha^{MB})^2; R_3, \alpha^{MB}(1 - \alpha^{MB}); R_4, \alpha^{MB}).
\end{aligned} \tag{18}$$

The above approach is motivated by research in decision neuroscience that adopts a similar model-based framework (Glascher et al., 2010; Lee, Shimojo, and O'Doherty, 2014;

and Dunne et al., 2016). Just as there is evidence that the brain encodes reward prediction errors, so there is evidence that it encodes state prediction errors analogous to the one in square brackets in (17) (Glascher et al., 2010).[12]

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. We extend the above model-based algorithm in a similar way, allowing for different model-based learning rates, $\alpha_+^{MB}$ and $\alpha_-^{MB}$, depending on whether the latest net stock market return is positive or negative. Specifically, at time $t$,

$$p^{\text{new}}(R_m = R) = \alpha_{t,+}^{MB} \text{ for } R \geq 1, \tag{19}$$

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

$$p^{\text{new}}(R_m = R) = \alpha_{t,-}^{MB} \text{ for } R < 1, \tag{20}$$

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

With this perceived return distribution in hand, how does the individual come up with an estimate of $Q^*(a)$, the value of choosing an allocation $a$ on some date and then continuing optimally thereafter? Once again, we follow an approach taken by experimental studies in decision neuroscience (Glascher at al., 2010). We assume that the individual estimates $Q^*(a)$ at time $t$ by taking equation (14) for the correct value of $Q^*(a)$ and applying it for his *perceived* time $t$ return distribution:

$$Q_t^{MB}(a) = E_t^p \log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E_t^p \log((1-a^*)R_f + a^* R_{m,t+1}), \tag{21}$$

where

$$a^* = \arg\max_a E_t^p \log((1-a)R_f + aR_{m,t+1}) \tag{22}$$

and where (21) differs from (14) only in that the expectation $E$ under the correct distribution has been replaced by the expectation under the investor's perceived distribution at time $t$, $E_t^p$.

[12]Our model-based algorithm shares the style of almost all economic frameworks in that the individual makes decisions with the help of a model of the environment; indeed, it is similar to one specific 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.

## 2.4 A hybrid model

An influential framework in decision neuroscience posits that people make decisions using a combination of model-free and model-based systems (Glascher et al., 2010; Daw et al., 2011). Specifically, it proposes that, at each time $t$ and for each possible action $a$, the individual computes a "hybrid" value of $Q(a)$ that is a weighted average of the model-free and model-based $Q$ values:

$$Q_t^{HYB}(a) = (1-w)Q_t^{MF}(a) + wQ_t^{MB}(a), \tag{23}$$

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:

$$p(a_t = a) = \frac{\exp[\beta Q_t^{HYB}(a)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(a')]}. \tag{24}$$

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

The model-free and model-based systems differ most fundamentally in how they estimate the value of an action: one system uses a model of the environment, while the other does not. However, there is another difference between them: the model-free system learns only from experienced returns, while the model-based system learns from all observed returns. 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 starts making decisions at time 0, the investor can look at historical charts and observe earlier stock market returns, which the model-based system can then learn from. To incorporate this, we extend the timeline of our framework so that it starts not at time 0 but $L$ dates earlier, at time $t = -L$. While the model-free system starts operating at time 0, the model-based system starts operating at time $-L$: it observes the $L$ stock market returns prior to time 0, $(R_{-L+1}, \ldots, R_0)$; uses these to form a perceived distribution of market returns as described in (19) and (20); and computes $Q^{MB}$ values based on that distribution, as in (21).[13]

[13] Having the model-free and model-based algorithms apply over different time intervals is not necessary for most of our applications in Section 4 – but it is important for one application, discussed in Section 4.2. Our implementation here is consistent with evidence in decision neuroscience. For example, Dunne

# 3 Model Properties

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

We use the timeline previewed at the end of the previous section. There are $L+T+1$ dates, $t = -L, \ldots, -1, 0, 1, \ldots, T$. Investors begin actively participating in financial markets at time 0. Their model-free systems therefore start operating only at time 0. However, before they begin their active participation, investors can use historical charts to observe stock market returns going back to time $t = -L$. Their model-based systems are therefore in operation 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.[14,15]

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 decision problem in (10) as efficiently as possible – then the answer would be to use learning rates that decline over time. Specifically,

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

[14]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. An alternative interpretation is that it is an approximation of a higher-frequency implementation. We have studied the effect of changing the model frequency. If we fix the learning rates $\alpha^{MB}$ and $\alpha^{MF}$ 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.

[15]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.

the time $t$ model-based learning rates in (19) and (20) would be[16]

$$\alpha_{t,+}^{MB} = \alpha_{t,-}^{MB} = 1/(t+1), \tag{25}$$

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

$$\sum_{t=0}^{\infty} \alpha_{t,\pm}^{MF} 1_{\{a_t=a\}} = \infty \qquad \sum_{t=0}^{\infty} (\alpha_{t,\pm}^{MF})^2 1_{\{a_t=a\}} < \infty, \tag{26}$$

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

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

## 3.1 An example

To show how the model-free and model-based systems work, we start with an example. We consider an investor who is exposed to a sequence of stock market returns from $t = -L$ to $t = T$, where $L = T = 30$. The returns are simulated from the distribution in (9) with $\mu = 0.01$ and $\sigma = 0.2$; these values provide an approximate fit to historical annual U.S. stock market data. We set the investor's learning rates to $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.3$, the exploration parameter $\beta$ to 50, the discount rate $\gamma$ to 0.98, and the degree of generalization $b$ to 0.0577. At each time, we allow the investor to choose his stock market allocation $a_t$ from one of 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$. We later examine how the coarseness of the action set affects the results.

As described in Section 2.4, the investors in our framework make decisions according to hybrid $Q$ values that combine the influences of the model-free and model-based systems. To clearly illustrate the mechanics of each system, we start by considering two simpler cases:

[16] Equation (25) assumes $L = 0$. For $L > 0$, the learning rates are $\alpha_{t,+}^{MB} = \alpha_{t,-}^{MB} = 1/(L+T+1)$.

one where the investor uses only the model-free system to make decisions, and one where he uses only the model-based system.

Table 1 shows the model-free $Q$ values, $Q^{MF}$, based on equations (13), (15), and (16) (upper panel) and the model-based $Q$ values, $Q^{MB}$, based on equation (21) (lower panel) that the investor assigns to each of the 11 allocation strategies on each of 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 $-0.045$ in bold at date 1 in the upper table indicates that the investor chose an allocation of 80% at date 0.[17]

Consider the upper panel of Table 1. The model-free system begins operating at time 0. At that time, then, it assigns a $Q$ value of zero to all the allocations. It then randomly selects the allocation 80%. 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 80% allocation therefore falls below zero. As per equations (15) and (16), the algorithm also engages in some generalization: since a 70% allocation and a 90% allocation are similar to an 80% 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 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 again chooses the allocation 30%. Since the market falls slightly at time 3, the time-3 $Q$ value of the 30% allocation goes down by a small amount. At dates 3 and 4, the investor chooses allocations of 50% and 40%, respectively, and updates the values of these allocations and their close neighbors based on the prediction errors they lead to at dates 4 and 5.

The lower panel shows that the $Q$ values generated by the model-based system are quite different. By time 0, the model-based system has already been operating for 30 periods and so already has well-developed $Q$ values for each of the 11 allocation strategies. In the periods

[17]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 (13). In our setting, for the model-based system, this probabilistic choice does not offer the usual exploration benefits: in each period, the investor learns the same thing about the distribution of stock market returns regardless of which allocation he chooses. We keep the probabilistic choice to allow for a more direct comparison with the model-free system. For the same reason, whenever we consider the model-based system in isolation, we allow for exploration, unless otherwise specified.

immediately preceding time 0, the simulated stock market returns are somewhat negative; lower 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 all the more: the negative stock market return at time 1 makes the investor's perceived distribution of stock market returns less appealing; this has a larger impact on portfolio strategies that allocate more to the stock market. At time 2, the stock market return is positive, so all $Q$ values go up, but those of the riskier allocations do so all the more.

Table 1 makes clear a key difference between the model-free and model-based systems: while, at each time, the model-based system updates the $Q$ values of all the allocations, the model-free system updates only the $Q$ values of the most recently chosen allocation and those of its nearest neighbors. The reason is that it is model-free: it knows nothing about the structure of the problem and therefore cannot make a strong inference, after seeing the outcome of an 80% allocation, about the value of a 20% allocation. On the one hand, this feature of the model-free system is what makes it a powerful and indispensable tool for both humans and animals: because it does not need any knowledge of the environment, it can be used in any setting. On the other hand, in the context of a portfolio problem, which has a strong underlying structure, the model-free approach can be inefficient.

## 3.2 Dependence on past returns

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

To study this, we take $300,000$ investors and expose each of them to a different sequence of simulated stock market returns from $t = -L$ to $t = T$. We then take investors' final stock market allocations $a_T$ at time $T$, regress them on the past 30 annual stock market returns $\{R_{m,T},\, R_{m,T-1},\ldots,R_{m,T-29}\}$ the investors have been exposed to, and record the coefficients. We do this for three cases, namely those where investor allocations are determined by the model-free system alone; by the model-based system alone; and by the hybrid system. In each case, we take $L = 30$ and $T = 30$. All investors have the same constant learning rates $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.3$. As before, we take $\beta = 50$, $\gamma = 0.98$, $\mu = 0.01$, and $\sigma = 0.2$. For ease of interpretation, we turn off generalization for now, so that $b = 0$.[18] Finally, we set $w = 0.5$,

[18]We use "$b = 0$" as shorthand for model-free learning without generalization. When $b = 0$, we compute

so that the hybrid system puts equal weight on the model-free and model-based systems. We will later look at how changing the values of key model parameters affects our results.[19]

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

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

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

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

model-free $Q$ values using equation (12) rather than equations (15)-(16), although the latter equations give the same result as $b \to 0$.

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

is stronger in the case of the 80% allocation: a high stock market return leads to a larger reward prediction error for the 80% allocation than for the 20% allocation, and hence a larger increase in the $Q$ value of the 80% allocation than of the 20% allocation. Given that this mechanism is less direct than the one for the model-based system, it is natural that, as shown in Figure 1, the dependence of the allocation on recent stock market returns is quantitatively smaller for the model-free system.

The weights that the model-based system puts on past returns decline as we go further into the past. Mathematically, this is because every time the model-based system updates its perceived return distribution, it downweights past returns by a proportional factor, 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 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 allocations put 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? Whenever the model-free system updates the $Q$ value of an action, this downweights the influence of past returns on this $Q$ value, relative to the most recent return. However, this effect passes through to allocation choice in a much more gradual way than for the model-based system because, at each time, the model-free system primarily updates only one $Q$ value; in short, it learns slowly. The bottom graph in Figure 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 changing the learning rates does not much affect the slope of the relationship; in all cases, the weights on past returns decline gradually as we go further into the past.

For many parameter values, including those used in Figures 1 and 2, the model-free allocation puts more weight on recent than distant past returns. However, the model-free system can exhibit much richer behavior than this. For example, it sometimes puts more weight on distant than on recent past returns. Moreover, the relationship between allocations and past returns is affected by factors that play no role for the model-based system.

To illustrate this, the four graphs in Figure 3 vary some key model parameters while keeping the others fixed at the benchmark levels listed above. The top-left graph in Figure 3 plots the coefficients in a regression of the model-free allocation on past stock market returns for four values of the generalization parameter $b$: 0, 0.0577, 0.115, and 0.23. The first of these values corresponds to the model-free system without generalization; the other three values give the Gaussian function in (16), normalized as a probability distribution, a standard deviation equal to that of a uniform distribution with width 0.2, 0.4, and 0.8, respectively.

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

The top-right graph in Figure 3 plots the relationship between the model-free allocation and past returns for three different values of $\beta$, which controls the degree of exploration, namely 10, 50, and 500. Recall that, as $\beta$ rises, the investor explores less; in other words, he is more likely to choose the allocation with the highest estimated $Q$ value. We find that, for a wide range of values of $\beta$ – any $\beta$ below 100 – the relationship between the allocation and past returns is qualitatively similar to that for our benchmark case of $\beta = 50$. However, when $\beta$ is very high – higher than 100 – we begin to see an increasing relationship between allocations and past returns. To see why, suppose that, at time 1, right after the investor enters financial markets, the stock market posts a high return, raising the $Q$ value of his initial 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 an outsize effect on his subsequent allocations.

The bottom-left graph plots the relationship between the model-free allocation and past returns for three different values of the discount rate $\gamma$, namely 0.3, 0.9, and 0.99. We find that, as we lower $\gamma$, the allocation puts much greater weight on recent past returns. This is a striking result that stands in contrast to the model-based system, where the discount rate does not affect the dependence of allocations on past returns.

Thus far, we have allowed investors to select from one of 11 possible allocations. The bottom-right graph in Figure 3 shows how the time $T$ allocation depends on past returns as we vary the number of allocation options available to investors, ranging from three, namely {0%, 50%, 100%}, up to 21, namely {0%, 5%, ... , 95%, 100%}. The graph shows something striking: as we lower the number of possible allocations, the relationship between the time $T$

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

One way of understanding the bottom-right graph is to note that reducing the number of allocation options is akin to increasing the degree of generalization: since generalization leads the investor to treat nearby allocations in a similar way, a large number of allocations coupled with generalization is like a small number of allocations without generalization. Just as in the top-left graph we see an increasing relationship between the time $T$ allocation and returns for high levels of generalization, so in the lower-right graph, we see an 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 returns. In some cases, this relationship is downward-sloping, and in others, upward-sloping. Moreover, the relationship between model-free allocations and past returns is affected by factors that play little to no role in the model-based system. We return to some of these novel implications in Section 4.

### 3.3 Dispersion and variability in allocations

We now consider some other properties of the model-free and model-based systems – properties related to the dispersion and variability in investor allocations. By "dispersion," we mean the standard deviation, across investors, of their date $T$ allocations. By "variability," we mean the standard deviation of investors' allocations over time: for each investor in turn, we compute the standard deviation of his allocations over time – the standard deviation of $\{a_{T-j}\}_{j=0}^{29}$ for this investor – and then average these standard deviations across investors. We obtain two results. The stronger result is that the variability in investor allocations is substantially lower under the model-free system: under this system, there is more "inertia" in an investor's allocations from period to period. The second result is that, under the model-free system, there is more dispersion in investors' final allocations.

To demonstrate these results, we now allow for dispersion in learning rates across investors.[20] Specifically, for each investor, we draw each of their learning rates – each of

[20]Data on investor beliefs about future stock market returns suggest that there is substantial dispersion in learning rates across investors. Giglio et al. (2021) analyze such data and find that an individual fixed effect explains more of the variation in beliefs than a time fixed effect: some investors are persistently optimistic

$\alpha_+^{MF}$, $\alpha_-^{MF}$, $\alpha_+^{MB}$, and $\alpha_-^{MB}$ – from a uniform distribution centered at $\bar{\alpha}$ and with width $\Delta$. Our benchmark parameter values are $L = T = 30$, $\bar{\alpha} = 0.3$, $\Delta = 0.3$, $\beta = 50$, $\gamma = 0.98$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$, so that there is some generalization. We take 10,000 investors, expose all of them to the same sequence of stock market returns from $t = -L$ to $t = T$, and compute the resulting dispersion and variability. We repeat this exercise 100 times for different return sequences and average the resulting set of dispersion and variability estimates.

The solid and dashed lines in the top three graphs in Figure 4 plot the variability of investor allocations under the model-based and model-free systems, respectively, as we vary three model parameters – the exploration parameter $\beta$, the mean learning rate $\bar{\alpha}$, and the dispersion $\Delta$ of learning rates – while keeping the other parameter values fixed at their benchmark levels. The main finding is that the dashed line is substantially below the solid line: the model-free system leads to lower variability than the model-based system. To understand this, note that, under the model-based system, investors tend to increase their allocation following a good return and lower their allocation following a poor return; as a consequence, there is substantial variability. By contrast, under the model-free system, an investor can become "stuck" at a particular allocation: if, early on, the investor chooses some allocation to the stock market and the market then performs well, the $Q$ value of the chosen allocation will be pushed up, raising the chance that the investor will keep choosing this allocation in subsequent years. The three upper graphs show that the difference in variability levels between the two systems is increasing in the mean learning rate and decreasing in the amount of exploration and the dispersion in learning rates.

The solid and dashed lines in the three lower graphs in Figure 4 plot the dispersion in final allocations across investors for the model-based and model-free systems, respectively, as we vary $\beta$, $\bar{\alpha}$, and $\Delta$, while keeping the other parameters fixed at their benchmark levels. The results for dispersion are not as clear cut as for variability. Nonetheless, the figures show that dispersion in allocations is typically higher for the model-free system. To understand this, note that, under the model-based system, following a high stock market return, all investors perceive an improvement in the distribution of stock market returns and hence tend to raise their allocation to the stock market; this, in turn, tends to keep the dispersion in allocations across investors at a relatively low level. The model-free system, by contrast, generates higher dispersion. This stems from the combination of the probabilistic action choice and the reinforcement inherent in this system. At time 0, the probabilistic action choice in (13)

while others are persistently pessimistic. Capturing this in our framework requires substantial dispersion in learning rates across investors, a claim we have confirmed in simulated data: as we increase this dispersion, individual fixed effects explain more of the variation in beliefs. Intuitively, investors with high $\alpha_+^{MB}$ and low $\alpha_-^{MB}$ are persistently optimistic, while those with low $\alpha_+^{MB}$ and high $\alpha_-^{MB}$ are persistently pessimistic.

leads to dispersed allocations across investors. If the stock market then performs well, this reinforces each investor's initial allocation, leading each investor to persist with his initial allocation and preserving the dispersion in allocations across investors.

In Section 4, we use the results of this section to shed light on some empirical facts about investor behavior.

### 3.4 Under- and over-reaction

It is natural to ask whether the model-based and model-free systems under- or over-react to news. We find that each of the two systems can generate both under- and over-reaction. However, relative to the model-based system, the model-free system exhibits a stronger degree of underreaction.

To study this, we define the reactivity of a system in a simple way: as the sensitivity of an investor's allocation to the latest stock market return. We compute this separately for the model-based and model-free systems and compare the results to a benchmark rational level of reactivity.

Specifically, we take 10,000 investors and expose each of them to a different sequence of stock market returns. At each time $t$, from $t = 1$ to $t = 60$, we run a cross-sectional regression of the change in investors' allocations $a_t - a_{t-1}$ on the time $t$ stock market return they were exposed to and record the slope coefficient. We set $L = 0$, $T = 60$, $\bar{\alpha} = 0.3$, $\Delta = 0.3$, $\gamma = 0.98$, $\beta = 50$, $b = 0.0577$, $\mu = 0.01$, and $\sigma = 0.2$. We do this analysis three times, for the cases where investors use the model-free system; the model-based system; and a model-based system with the declining learning rates in (25), which serves as the rational benchmark.[21]

Figure 5 presents the results. The solid line, dashed line, and dash-dot lines correspond to the model-based system, model-free system, and rational benchmark, respectively; each line plots the sensitivity of investor allocations to returns at each moment of time. The region below the dash-dot line represents underreaction and the region above it, overreaction. The graph shows that the model-based system underreacts in the early periods and overreacts in the later periods. This is because it uses a constant learning rate. In the early periods, when there is a lot to learn from each market return, the investor should be using a higher learning rate than average, as in (25); his constant learning rate therefore leads him to underreact.

[21] In the case of the rational benchmark, we turn off exploration, as it is rational for each investor to choose the allocation with the highest estimated $Q$ value at each time.

In later periods, there is less to learn from each market return; the investor should therefore be using a lower learning rate than average, as in (25); his constant learning rate leads him to overreact. Put differently, the model-based system overreacts most of the time because, by using a constant learning rate, it is effectively assuming that the environment is non-stationary, even though, given that returns are i.i.d., it is actually stationary.

The same considerations also apply to the model-free system: it too uses a constant learning rate. As such, it initially underreacts, but in the longer run – a longer run than is visible in the figure – it overreacts. However, the more salient result is that, relative to the model-based system, the model-free system underreacts: the dashed line is substantially below the solid line. This is because it learns slowly: at each time, it updates primarily the $Q$ value of the most recently chosen action. As a result, the set of 11 $Q$ values corresponding to the 11 actions do not change much from period to period.

# 4 Applications

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

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

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

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

We adopt the following simple parameterization. As in Section 3, we set $\mu = 0.01$ and $\sigma = 0.2$. Each investor is trying to solve the problem in (10) and chooses allocations according to the hybrid system in (23)-(24). As before, at each time, the investor chooses from the 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$. 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 $\Delta$. We set $\bar{\alpha} = 0.5$, $\Delta = 0.3$, $\beta = 50$, $\gamma = 0.98$, $b = 0.0577$, and $w = 0.5$, so that investors put equal weight on the model-free and model-based systems. Later, we will formally estimate the value of $w$ that best fits the data.

We now use the above structure to discuss a range of applications.

## 4.1 Extrapolative demand

The first application follows directly from the analysis of Section 3.2, but it is an important one that merits further discussion. A common assumption in psychology-based models of asset prices and investor behavior is that people have extrapolative demand: their demand for a financial asset depends positively on the asset's past returns, and especially on its recent past returns.[22]

The framework of Section 2 provides a new foundation for such extrapolative demand, one rooted in the model-free system. As shown in Section 3.2, for a wide range of parameter values, the model-free system generates an allocation that depends positively on past returns and more so on recent past returns – and it does so for a reason that is new to the finance literature. To summarize the mechanism from Section 3.2: following a high stock market return, the reward prediction error is higher if the investor has a high allocation to the stock market than if he has a low allocation; he is therefore more likely to choose a high allocation going forward.

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

The solid line in Figure 6 is similar to the line marked "Hybrid" in Figure 1 in that both lines correspond to decisions made under the hybrid system. However, the two lines differ in that, relative to the analysis of Section 3.2, we are now allowing for dispersion across investors in their learning rates and for multiple cohorts, and are using a higher mean learning rate. The multiple cohorts in particular make the solid line in Figure 6 decline more quickly than the "Hybrid" line in Figure 1: some of the investors present in the market at time $T = 30$ entered only at time 25; as such, their model-free system puts zero weight on returns before time 25.

The framework of Section 2 offers another insight relative to the existing finance literature on extrapolative demand, namely that this extrapolative demand has two different sources which operate on different time scales: one that stems from a model-based system that puts heavy weight on *recent* returns, and one that stems from a model-free system that puts substantial weight even on *distant* past returns; indeed, in some cases, the model-free system puts more weight on distant than on recent returns. As such, while the allocations of real-world investors appear to put more weight on recent returns, this may mask a model-free component that puts more weight on distant returns, but that is outweighed by a model-based component that puts heavy weight on recent returns.

[22] A very partial list of papers that study extrapolative demand is Cutler, Poterba, and Summers (1990), De Long et al. (1990), Barberis et al. (2015, 2018), Cassella and Gulen (2018), Jin and Sui (2021), Liao, Peng, and Zhu (2021), and Pan, Su, and Yu (2021).

## 4.2 Experience effects

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

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

There are two key features of experience effects that we hope to replicate. The more important one is that, if an investor begins participating in financial markets at time $t$, his allocation to the stock market should depend substantially more on the stock market return at time $t+1$, $R_{m,t+1}$ – a return he experienced – than on the stock market return at time $t$, $R_{m,t}$, a return he did not experience. Put differently, if we plot the coefficients in a regression of an investor's allocation on past returns, we should see a "jump" 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 an investor's allocation on the past stock market returns he has experienced should decline for more distant past returns. As a way of capturing both features, Malmendier and Nagel (2011) propose that investors' decisions are based on a weighted average of past returns in which, for an investor with $n$ years of experience, the weight on the return $k$ years ago is

$$(n-k)^{\lambda}/A, \tag{27}$$

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

To see whether our framework can generate the two features of experience effects, we proceed as follows. For each of the six cohorts, we take the $50,000$ investors in the cohort and regress their time $T$ allocations $a_T$ on the past 30 years of stock market returns. Figure 7 presents the results. The six graphs correspond to the six cohorts. In each graph, the solid line plots the coefficients in the above regression, normalized to sum to one so that we can properly compare them to the Malmendier and Nagel (2011) coefficients in (27). The dashed line plots the functional form in (27) for the cohort in question, and the vertical dotted line marks the point at which the cohort enters financial markets.

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

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

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

In Section 4.8, we estimate the parameter values – specifically, the mean model-free and model-based learning rates and the weight $w$ on the model-based system – that best fit the evidence on experience effects. The estimated $w$ is 0.5, which reflects an equal mix of the two systems. The intuition is that the model-free system explains why allocations depend even on experienced returns in the distant past, while the model-based system explains the substantially greater influence of recent returns.

## 4.3 Investor beliefs and the frequency disconnect

Several studies have found that investor beliefs about future stock market returns are a positive function of recent past stock market returns. Our framework can capture this; but more strikingly, it can also help explain two puzzling disconnects between investor beliefs and investor actions – one in the frequency domain, which we discuss in this section, and one in the cross-section of investors, which we discuss in the next section.

The disconnect in the frequency domain is simple to state. While studies of investor expectations about future returns find that these expectations depend heavily on *recent* past returns – see Greenwood and Shleifer (2014) – studies of investor stock market allocations find that these depend to a substantial extent even on distant past returns (Malmendier and Nagel, 2011).[23]

Two features of our framework allow it to explain this disconnect. First, while each investor's decision is based on both the model-free and model-based systems, only one of these – the model-based system – has an explicit role for investor beliefs. Second, the model-free system, which is unrelated to beliefs, recommends allocations that put substantially more weight on distant past returns than does the model-based system. Taken together, these features mean that the investor's beliefs, which are generated by the model-based system, will put heavy weight on recent returns, while his allocations, which are based on both systems, will put a greater relative weight on distant past returns. As such, the framework drives a wedge between actions and beliefs.

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

A number of studies find a positive time-series relationship between investor beliefs and allocations. For example, Greenwood and Shleifer (2014) find that the average investor expectation of future stock market returns is correlated with flows into equity market mutual funds. Our framework is consistent with these findings. In our simulated data, there is a strong time-series correlation between investor allocations and beliefs, both at the individual and aggregate levels. However, the model-free system dampens the economic magnitude of

[23] We can formalize this in the following way. When Malmendier and Nagel (2011) use the weights in (27) to characterize the relationship between an investor's allocation and the past returns he has experienced, they obtain an estimate of $\lambda \approx 1.5$. Suppose that we now take the functional form in (27) and use it, with $n = 30$, to characterize the relationship between investor *beliefs* and the past 30 years of stock market returns. Using Gallup data on stock market expectations from October 1996 to November 2011, we find that the best fit is for $\lambda \approx 50$, which puts a much greater weight on recent returns.

this relationship because it is less reactive to recent returns.

## 4.4 Investor beliefs and the cross-sectional disconnect

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

Our framework can help capture this disconnect. The reason is that the model-based system, which generates an investor's beliefs, is only part of what drives his allocation decision; the model-free system, which is independent of beliefs, also has a large impact on this decision. To see the implications of this, suppose that the stock market posts a high return. The investor's expectation about the future stock market return will then go up significantly: the model-based system, which determines beliefs, puts substantial weight on recent returns. However, the investor's allocation will be less sensitive to the recent return: it is determined in part by the model-free system, which puts substantial weight even on distant past returns and less weight on the most recent return.

We now examine this quantitatively. The right-most column in Table 2 reports, for three different parameterizations, the coefficient in a regression of investors' allocations $a_T$ to the stock market at time $T$ on their expected returns on the stock market over the next year. More precisely, we take the 300,000 investors and expose each of them to the same sequence of stock market returns from $t = -L$ to $t = T$. We then run the regression of allocations on beliefs. We repeat this several times for different return sequences. The table reports the average regression coefficient across these multiple trials. The three rows of the table correspond to different values of $w$, the weight on the model-based system.

The table shows that our framework can help explain the cross-sectional disconnect described above: for our benchmark value of $w = 0.5$, the regression coefficient in our simulated data, 1.33, is similar to that obtained by Giglio et al. (2021) in actual data. Moreover, the table shows that the model-free system plays a key role in this result: as we increase the weight on the model-free system, the sensitivity of allocations to beliefs falls.

## 4.5 Dispersion and inertia in household allocations

Households differ in their asset allocation decisions: some participate in the stock market, while others do not; and among households that participate, the fraction of wealth they invest in the stock market varies substantially. It is not easy to explain these differing allocations: regressions of allocations on explanatory variables have a low $R^2$.

The framework of this paper offers two new ways of thinking about this dispersion in holdings. First, it says that these differences are due in part to differences across investors in their learning rates, namely $\alpha_+^{MF}$, $\alpha_-^{MF}$, $\alpha_+^{MB}$, and $\alpha_-^{MB}$. To examine this, we take the $50,000$ investors in cohort 1, expose them to the same simulated sequence of stock market returns between time $-L$ and time $T$; regress their final allocations $a_T$ at time $T$ on the differences in their model-free learning rates $\alpha_+^{MF} - \alpha_-^{MF}$ and the differences in their model-based learning rates $\alpha_+^{MB} - \alpha_-^{MB}$; and record the $R^2$. We repeat this exercise for many return sequences and compute the average $R^2$ across the different return sequences. We find it to be 7%, a substantial $R^2$ relative to existing predictors of allocations.

The second possibility is one discussed earlier in the paper in connection with Figure 4. The lower-right panel in that figure shows that the model-free system can generate substantial dispersion in investor allocations $a_T$ at time $T$ even when all investors have the *same* learning rates. The dispersion here is driven by the interaction of the probabilistic action choice and model-free reinforcement. If, as a result of the probabilistic choice, investor A chooses a low allocation to the stock market early on while investor B chooses a high allocation, and the stock market then posts a high return, choosing a low (high) allocation will be reinforced for investor A (B), leading to persistent differences in the investors' allocations.

There is substantial cross-sectional dispersion in households' allocations to the stock market – but there is also significant individual-level inertia in these allocations over time. This inertia is often attributed to transaction costs, procrastination, or inattention.

The framework in this paper offers a new way of thinking about inertia in investor holdings: it says that the inertia arises endogenously from the model-free system. In Section 3.3, and specifically in the upper panel of Figure 4, we see that, relative to the model-based system, the model-free system generates lower variability, or equivalently, higher inertia. If, after an investor chooses some allocation to the stock market, the market posts a good return, the $Q$ value of that allocation goes up substantially, which makes it more likely that the investor will keep choosing that allocation in the future.

## 4.6 Non-participation

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

A long-standing question asks why many U.S. households do not participate in the stock market; the traditional Expected Utility model, by contrast, predicts that all investors will allocate at least some fraction of their wealth to the stock market. Our framework can shed light on this. In particular, the model-free system tilts investors toward not participating. To see why, consider an investor who makes decisions according to the model-free system. If he allocates some money to the stock market but then experiences a poor market return, this raises the probability that, in a subsequent period, he will switch to a 0% allocation to the market. Importantly, once he does so, the model-free system will update only the $Q$ value of the 0% allocation: since, generalization aside, it learns only about the action taken, it stops learning about the stock market, and, in particular, fails to learn that the stock market has better properties than indicated by the poor return the investor experienced. This will tend to keep the investor at the 0% allocation for an extended period of time.

We illustrate this in a modified version of our framework with just two allocations: 0% and 100%. It is natural to use a two-allocation framework for this application because the participation decision has a binary flavor: Should I participate or not? It is not important that the stock market allocation is a 100% allocation; we obtain similar results if the two allocations are 0% and 50%, say.[24]

We take $10,000$ investors and expose each of them to a different sequence of stock market returns. For each investor, we compute the fraction of time between dates 0 and $T$ that he chooses a 0% allocation. In addition, for each investor, we identify the episodes where he allocates 0% to the stock market for multiple consecutive years and record the duration of the longest such episode. We do this exercise twice: first for the case where decisions are made by the model-free system, and then for the case where they are made by the model-based system. The parameter values are $L = T = 30$, $\bar{\alpha} = 0.3$, $\Delta = 0.3$, $\beta = 50$, $\gamma = 0.98$, $b = 0.0577$, $\mu = 0.01$, and $\sigma = 0.2$.

The results confirm that the model-free system tilts investors toward non-participation. Under the model-free system, 45% of investors spend more than 80% of the 30 years not participating in the stock market, in other words, at a 0% allocation. By contrast, under

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

the model-based system, just 3% of investors spend more than 80% of the 30 years not participating. In a similar vein, under the model-free system, 60% of investors have a non-participation streak that is at least 10 years long; under the model-based system, only 16% of investors have a streak of this length.

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

## 4.7 Persistent investment mistakes

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

To demonstrate this quantitatively, it is natural to consider a slightly different setting from the one we have used so far. In this new setting, there are ten risky assets. The gross return on asset $i$, $R_i$, is distributed as

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

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

We take as a rational benchmark the case where all investors use the model-based system with the declining learning rates in (25); for these learning rates, consistent with the i.i.d. assumption, investors are equally weighting the past returns on each asset. There are $10,000$ investors; we also set $L = 0$, $T = 30$, $\gamma = 0.98$, $b = 0.0577$, and $\beta = \infty$ so that there is no exploration. We find that, in this case, at time $T = 30$, 46% of investors are allocating to the best asset, asset 10.

We now consider the case where all investors instead use the model-free system with constant learning rates to tackle this problem. For each investor, their learning rates are drawn from a uniform distribution with mean $\bar{\alpha} = 0.3$ and width $\Delta = 0.3$; the exploration parameter is $\beta = 50$. In this case, we find that, at time $T = 30$, just 21% of investors are allocating to asset 10. Consistent with our claim above, then, the model-free system learns slowly: it takes longer to figure out the sensible course of action. This result does not hinge on the constant learning rate. If investors instead use the model-free system in conjunction with a declining learning rate – one that satisfies the conditions for long-run convergence of $Q$ values in (26) – then, at time $T = 30$, just 19% of investors are allocating to asset 10.[25]

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

### 4.8 Parameter estimation

Throughout this section, we have taken a simple parameterization of our framework and shown that it can provide a qualitative and approximate quantitative match to a number of facts about investor behavior. We now conduct a simple estimation exercise to see which parameter values best match the data. The parameters we estimate are the mean model-based learning rate across investors $\bar{\alpha}^{MB}$; the mean model-free learning rate $\bar{\alpha}^{MF}$; the exploration parameter $\beta$; and most important, the weight $w$ on the model-based system. We do the estimation in two steps. First, we use data on investor beliefs to estimate $\bar{\alpha}^{MB}$. With this in hand, we then estimate $\bar{\alpha}^{MF}$, $\beta$, and $w$ by targeting the experience effects in Malmendier and Nagel (2011) and the sensitivity of allocations to beliefs in Giglio et al. (2021). We set the remaining parameters to $L = T = 30$, $\gamma = 0.98$, $\Delta = 0.5$, $b = 0$, $\mu = 0.01$, and $\sigma = 0.2$.

We estimate the mean model-based learning rate by searching for the value of $\bar{\alpha}_{MB}$ that best fits the empirical relationship between investor beliefs and past returns. We take Gallup data, from October 1996 to November 2011, on beliefs about future stock market returns and regress these beliefs on past annual stock market returns. The coefficient on the past year's return is 0.127, and the coefficient on the return two years in the past is 0.037; the ratio of the two coefficients is 0.29. We search for a value of $\bar{\alpha}_{MB}$ that, in simulated data, best matches the first coefficient, 0.127, and the rate of decline in the coefficients, 0.29;

[25] Specifically, we use the learning rate $1/(1 + t^{0.6})$, which satisfies the conditions in (26).

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

$$(c_1 - 0.127)^2 + (\frac{c_2}{c_1} - 0.29)^2. \tag{28}$$

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

With this value of $\bar{\alpha}^{MB}$ in hand, we now estimate $\bar{\alpha}^{MF}$, $\beta$, and $w$. To do so, we target two quantities. The first is the coefficient in a regression of investor allocations on investor beliefs; Giglio et al. (2021) find this coefficient to be approximately 1 in the data. For given values of $\bar{\alpha}^{MF}$, $\beta$, and $w$, we can compute this coefficient, $d$, in our simulated data. Our second target is the functional form in (27) which Malmendier and Nagel (2011) use to capture empirical experience effects; intuitively, we are looking for parameter values that minimize the distance between the red and blue lines in the six graphs in Figure 7.

We search for values of $\{\bar{\alpha}^{MF}, \beta, w\}$ that best match the two targets. Specifically, for cohort 1, for given values of $\bar{\alpha}^{MF}$, $\beta$, and $w$, we run a regression, in our simulated data, of the time $T$ allocations on the past 30 years of returns and compute the $L^2$ norm of the difference between the values of the 30 coefficients (the blue line in the top-left graph in Figure 7) and the 30 values implied by (27) (the red line in the graph). We call this $\text{MSE}_1$. In a similar way, we compute $\text{MSE}_i$ for $i = 2$ to 6, which correspond to the other five cohorts. We repeat this exercise for many different values of $\{\bar{\alpha}^{MF}, \beta, w\}$ and identify the values of these parameters that minimize

$$\sum_{i=1}^{6} \text{MSE}_i + (d - 1)^2. \tag{29}$$

The first term in (29) targets empirical experience effects, while the second term targets the empirical sensitivity of allocations to beliefs.

We obtain parameter values of $\bar{\alpha}^{MF} = 0.7$, $\beta = 20$, and $w = 0.5$. The first two parameters are not precisely identified. This can be understood by looking at Figures 2 and 3: as we vary $\bar{\alpha}^{MF}$ and $\beta$, the dependence of allocations on past returns does not vary strongly, making it harder to estimate these parameters.

The value of $w$, by contrast, *is* well identified. The reason is the following. In the first term in (29), we are trying to match the empirical pattern of experience effects, as summarized by the functional form in (27). As shown by the red lines in Figure 7, this involves both an initial sharp decline in the coefficients on past returns, but also a significant dependence on distant past experienced returns. As shown in the upper panel of Figure 2, the model-based system can capture the initial sharp decline in coefficients, but, when calibrated to do so, it cannot capture the dependence on distant past returns. By contrast, the lower panel of Figure 2 shows that the model-free system can capture a high dependence on distant past returns but not the initial sharp decline. As such, to match both features of the data, we need to put substantial weight on both systems – as it turns out, a roughly equal weight on the two systems.

## 4.9 Extensions

We now discuss some possible extensions of our framework.

**Time-varying learning rates.** We have taken each investor's learning rates to be constant over time; even this simple case has many applications. Nonetheless, learning rates may vary over time. For example, there is evidence that they go up at times of greater volatility or dramatic news. Such an assumption can be incorporated into our framework and may lead to new predictions.

**Time-varying weights on the two systems.** We have taken $w$, the weight on the model-based system, to be constant over time. A well-known hypothesis in psychology proposes that $w$ varies over time (Daw, Niv, and Dayan, 2005): the individual puts more weight on the system that is currently more certain about the values of different courses of action. For example, in the early stage of a person's interaction with a new environment, it may take a high value: the model-based system learns quickly and is therefore more useful. Over time, as the model-free system accumulates more experience, the individual may start to put more weight on it, lowering $w$. In our framework, this would predict that older people display a lower degree of extrapolation – in other words, react less to recent returns – and exhibit more inertia in their portfolio holdings.

**Other model-based frameworks.** When we specify the model-free system in Section 2, we do not have much flexibility. Model-free systems tend to behave similarly: the individual takes an action, and based on the outcome, he updates the value of the action. Indeed, when we replace with Q-learning with SARSA, an alternative model-free framework, we obtain similar results. However, when specifying the model-based part of our framework, we have a wider range of choices. In Section 2, we adopted a model-based system inspired by those used in psychology, but others are possible. For example, some investors may use a model-based system with a more contrarian flavor – one that, following a good return, recommends a lower allocation on the grounds that the market may now be overvalued. When incorporated into our framework, this model-based framework can lead to a tension between the model-free and model-based systems whereby, after a good return, the model-free system wants to increase exposure to the stock market while the model-based system wants to reduce it.

**State dependence.** Thus far, we have not allowed for state dependence: we consider action values $Q(a)$ rather than action-state values $Q(s, a)$; even this simple case has many applications. However, both the model-based and model-free systems are capable of handling state dependence. One possible state variable $s_t$ is the valuation of the stock market – its $P/E$ ratio, for example. In this case, the model-based and model-free frameworks may make opposite recommendations. Since a high market valuation is likely preceded by high returns, the model-free system may recommend a higher allocation to the stock market, as discussed in Section 3.2. However, the model-based system may recommend a lower allocation on the grounds that a market with a high valuation may be overvalued. Once again, the two systems pull the investor in different directions.

We cannot study this idea in the partial equilibrium framework of this paper: a single investor cannot impact the market valuation. However, in an equilibrium framework, investors' collective actions affect the market's value, allowing for an examination of the above argument.

**Inferring beliefs from the model-free system.** Until now, we have associated beliefs with the model-based system. However, it is possible that investors also use the model-*free* system to make inferences about beliefs. When an investor is surveyed about his beliefs – for example, his beliefs about the stock market's future return or risk – it is natural that he will first consult the model-based system, which will give him a direct measure of beliefs. However, he may also consult the model-free system, and if he finds that $Q^{MF}(a = 1) > Q^{MF}(a = 0)$, so that the model-free system judges the stock market to have better prospects than the risk-free asset, he may take this as a signal that the stock market has better *properties* – for example, a higher expected return and lower risk.

This idea can help us make sense of some puzzling facts. For example, Giglio et al. (2021) find that, when investors expect high returns in the stock market, they also expect the market to have lower risk. Our framework explains this in the following way. When investors are surveyed for their expectations about returns or risk, they naturally first consult their model-based system, which gives them direct estimates of return and risk. However, they may also consult their model-free system, and if they find that $Q^{MF}(1) > Q^{MF}(0)$, so that the stock market has better prospects than the risk-free asset, they may take this as a signal that the stock market has better properties – for example, both a higher expected return and lower risk. Conversely, if they consult the model-free system and find that $Q^{MF}(1) < Q^{MF}(0)$, so that the stock market has worse prospects, they may take this as a signal that the stock market has worse properties – for example a lower expected return and higher risk. This may explain why investors perceive a negative relationship between risk and return.

# 5 Discussion

In Section 4, we saw that the framework of Section 2 can shed light on a range of facts about investor behavior. Beyond this, our analysis also points to some broader themes.

First, as shown in Section 4.8, the parameters of the framework that best fit the data put substantial weight on the model-free system. It is striking that investors would put weight on a system that uses no information about the probabilistic structure of returns, even though such information is in principle available. This may be an indication that many households have a poor sense of the structure of asset returns and therefore fall back on a system – the model-free system – that does not require any knowledge of this structure. It may also be a sign of how fundamental the model-free system is to human decision-making: after all, it likely played a role in guiding human behavior through much of evolutionary history.

Second, when researchers try to explain the variation, across investors, in their allocations to the stock market, they focus on differences in beliefs about future returns or risk, or on differences in risk aversion. The framework of this paper suggests that learning rates, both model-free and model-based, are a deeper driver of the variation in risky holdings. One goal for future research may therefore be to estimate these learning rates at the individual level and to connect them to financial holdings.

Third, in most models of investor behavior, the primitives are investors' beliefs and preferences. These then lead, by way of a Bellman equation, to a value function. The model-based system in our framework also has this feature. However, the model-free system is different: here, it is the value function itself that is the primitive, the object that the investor updates from period to period. As described in Section 4.9, the investor may use the value function to draw inferences about the appropriate beliefs to hold about the risky asset; as such, the value function may generate beliefs, rather than the other way round.

Finally, a large number of papers in economics take a dynamic investment problem, use mathematical or numerical techniques to derive the value function, and then use this value function to interpret observed behavior. However, this line of research rarely explains how an individual might actually come to act in the way described by the value function – a natural question to ask, given that few people know how to solve Bellman equations. By contrast, in this paper, we try to make sense of individual behavior using a framework that is rooted in algorithms that the brain is thought to use when estimating the value of a course of action.

# 6 Conclusion

In the past decade, researchers in psychology and neuroscience studying human decision-making have increasingly adopted a framework that combines two systems, namely "model-free" and "model-based" learning. We import this framework into a simple financial setting, study its properties, and link it to a range of applications. We show that it provides a foundation for extrapolative demand and experience effects; resolves a puzzling disconnect between investor allocations and beliefs in both the frequency domain and the cross-section; can help explain the dispersion across investors in their stock market allocations as well as the inertia in these allocations over time; and can shed light on why many households make persistent investment mistakes. More broadly, the framework offers a way of thinking about individual behavior that is grounded in recent evidence on the computations that the brain undertakes when estimating the value of a course of action.

# 7 Appendix

**A portfolio-choice problem that fits the model-free / model-based learning framework (Section 2.2)**

The investor's objective is to maximize

$$(1-\gamma)E(\log W_1) + \gamma(1-\gamma)E(\log W_2) + \gamma^2(1-\gamma)E(\log W_3) + \ldots \tag{30}$$

where

$$W_t = W_0 \Pi_{\tau=1}^{t} R_{p,\tau} \tag{31}$$

is his wealth at time $t$. Substituting (31) into (30) and rearranging, the objective function becomes

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

as in (10).

# 8 References


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

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

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

Barberis, N. and R. Thaler (2003), "A Survey of Behavioral Finance," in G. Constantinides, M. Harris, and R. Stulz (Eds.), *Handbook of the Economics of Finance*, North Holland.

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

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

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

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

Collins, A. (2018), "Learning Structures Through Reinforcement," in *Goal-directed Decision-*

*making: Computations and Neural Circuits*, Academic Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

*Economics*, forthcoming.

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

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

Liao, J., Peng, C., and N. Zhu (2021), "Extrapolative Bubbles and Trading Volume," *Review of Financial Studies*, forthcoming.

Malmendier, U. and S. Nagel (2012), "Depression Babies: Do Macroeconomic Experiences Affect Risk-taking?" *Quarterly Journal of Economics* 126, 373-416.

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

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

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

Pan, W., Su, Z., and J. Yu (2021), "Extrapolative Market Demand," Working paper, Tsinghua University.

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

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

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

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

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

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

**Table 1. Model-free and model-based Q values.** The upper panel reports model-free Q values for 11 stock market allocations from $t = 0$ to $t = 5$. The lower panel reports model-based Q values for the 11 allocations for the same six dates. The rows labeled "net market return" report the net stock market return at each date. Boldface type indicates the allocation that was taken in the previous period. We set $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.3$, $\beta = 50$, $\gamma = 0.98$, $b = 0.0577$, $\mu = 0.01$, and $\sigma = 0.2$.

| MODEL-FREE | | | | | | |
|---|---|---|---|---|---|---|
| date | 0 | 1 | 2 | 3 | 4 | 5 |
| net market return | | -17.4% | 18.3% | -1.3% | 12.8% | -16.6% |
| 0% | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0 | 0 | 0.004 | 0.003 | 0.003 | 0.003 |
| 30% | 0 | 0 | **0.016** | **0.015** | 0.015 | 0.011 |
| 40% | 0 | 0 | 0.004 | 0.003 | 0.008 | **-0.008** |
| 50% | 0 | 0 | 0 | 0 | **0.023** | 0.019 |
| 60% | 0 | 0 | 0 | 0 | 0.005 | 0.005 |
| 70% | 0 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 |
| 80% | 0 | **-0.045** | -0.045 | -0.045 | -0.045 | -0.045 |
| 90% | 0 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 |
| 100% | 0 | 0 | 0 | 0 | 0 | 0 |

| MODEL-BASED | | | | | | |
|---|---|---|---|---|---|---|
| date | 0 | 1 | 2 | 3 | 4 | 5 |
| net market return | | -17.4% | 18.3% | -1.3% | 12.8% | -16.6% |
| 0% | 0.025 | 0 | 0.063 | 0.029 | 0.283 | 0 |
| 10% | 0.026 | -0.004 | 0.065 | 0.031 | 0.288 | **-0.002** |
| 20% | 0.027 | **-0.009** | **0.067** | 0.032 | 0.292 | -0.004 |
| 30% | 0.027 | -0.014 | 0.069 | 0.032 | 0.296 | -0.006 |
| 40% | 0.026 | -0.021 | 0.07 | 0.032 | 0.3 | -0.009 |
| 50% | 0.024 | -0.028 | 0.07 | 0.032 | 0.304 | -0.012 |
| 60% | 0.021 | -0.036 | 0.069 | **0.031** | **0.307** | -0.015 |
| 70% | 0.017 | -0.045 | 0.068 | 0.03 | 0.309 | -0.019 |
| 80% | 0.011 | -0.054 | 0.066 | 0.028 | 0.311 | -0.023 |
| 90% | 0.004 | -0.065 | 0.063 | 0.026 | 0.313 | -0.027 |
| 100% | -0.005 | -0.078 | 0.059 | 0.022 | 0.314 | -0.033 |

**Table 2. Sensitivity of investor allocations to investor beliefs.** The fourth column of the table reports the sensitivity of investors' stock market allocations $a_T$ at time $T$ to their time $T$ expectations of the future one-year stock market return for various values of the weight $w$ on the model-based system. There are 300,000 investors: six cohorts of 50,000 investors each which enter financial markets at different dates. We set $L = T = 30$, $\bar{\alpha} = 0.5$, $\Delta = 0.3$, $\beta = 50$, $\gamma = 0.98$, $b = 0.0577$, $\mu = 0.01$, and $\sigma = 0.2$.

| $\bar{\alpha}_{\pm}^{MF}$ | $\bar{\alpha}_{\pm}^{MB}$ | $w$ | Action vs. belief |
|---|---|---|---|
| 0.5 | 0.5 | 0.2 | 0.88 |
| 0.5 | 0.5 | 0.5 | 1.33 |
| 0.5 | 0.5 | 1 | 2.6 |

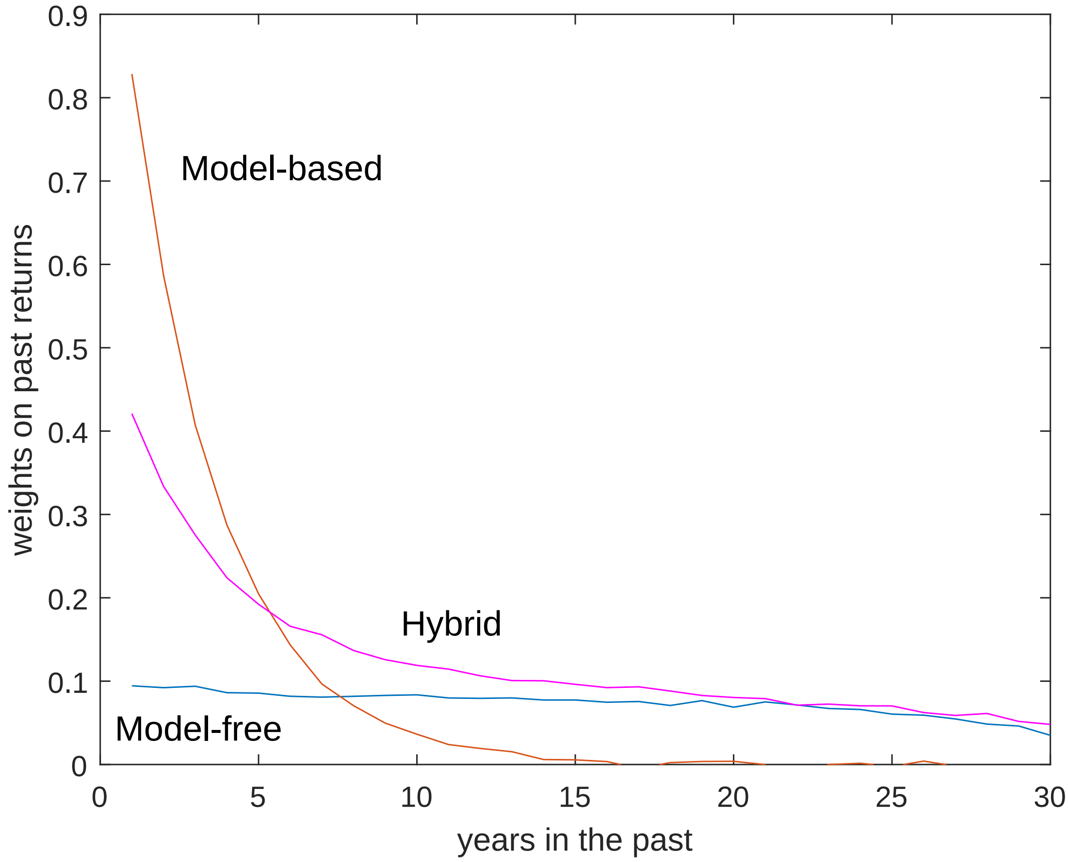


**Figure 1.** We run a regression of investors' allocations to the stock market $a_T$ at time $T$ on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors have been exposed to and plot the coefficients for three cases: a model-free system, a model-based system, and a hybrid system. There are 300,000 investors. We set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.3$, $\beta = 50$, $\gamma = 0.98$, $w = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

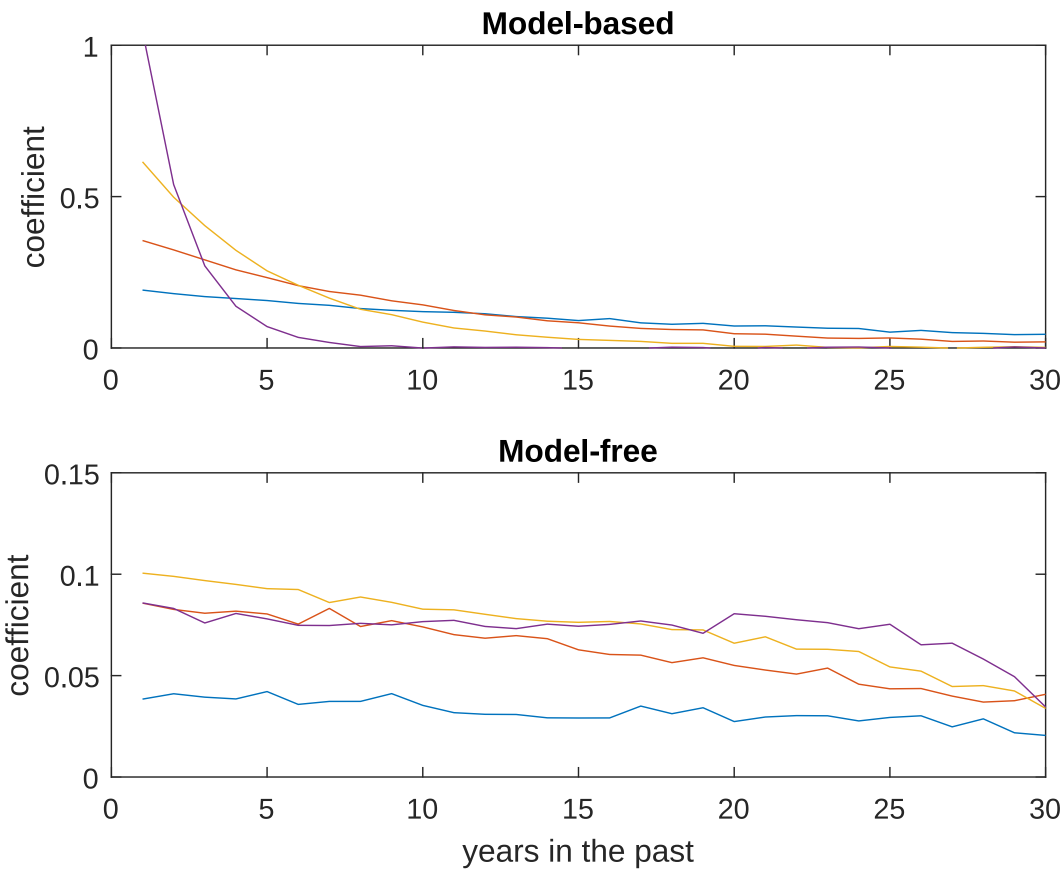


**Figure 2.** We run a regression of investors' allocations to the stock market $a_T$ at time $T$ on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors have been exposed to. The top graph plots the coefficients for the model-based system for four values of the learning rates $\alpha_{+}^{MB}$ and $\alpha_{-}^{MB}$, namely 0.05 (blue), 0.1 (red), 0.2 (yellow), and 0.5 (magenta). 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 (blue), 0.1 (red), 0.2 (yellow), and 0.5 (magenta). There are 300,000 investors. We set $L = T = 30$, $\beta = 50$, $\gamma = 0.98$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

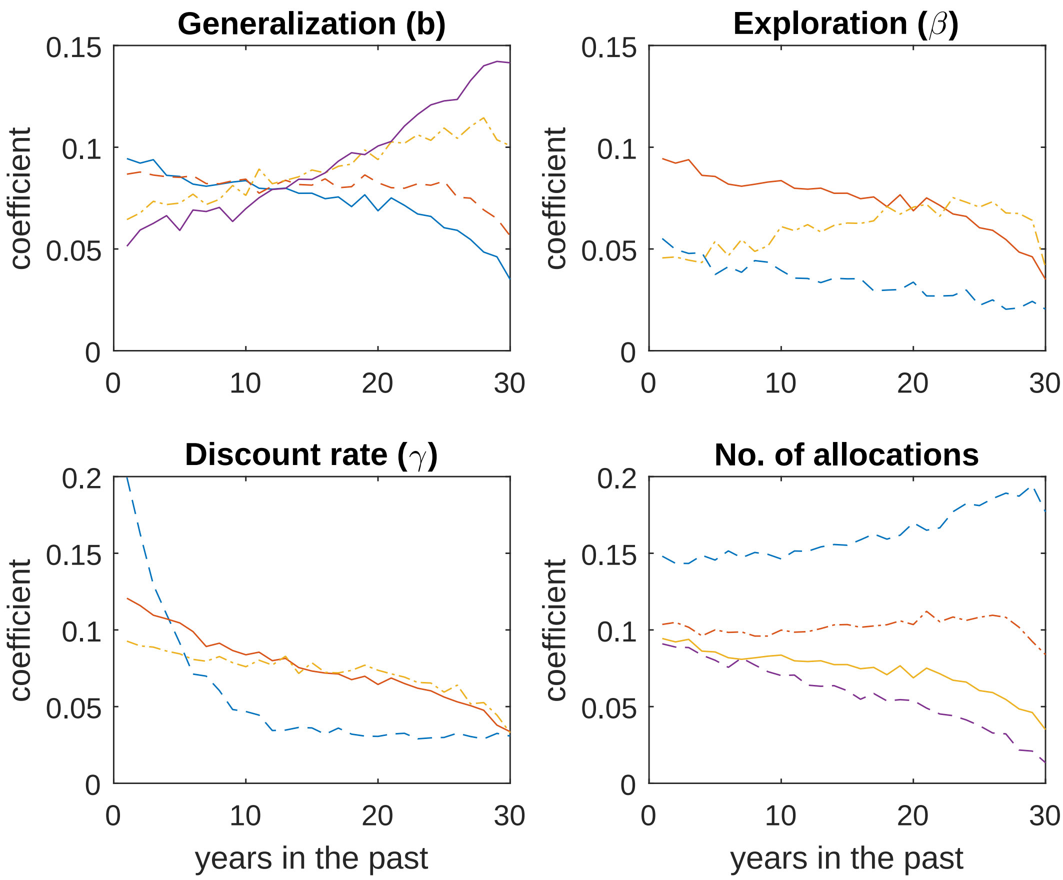


**Figure 3.** For different sets of parameter values, we run a regression of investors' allocations to the stock market $a_T$ at time $T$ under the model-free system on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ investors have been exposed to. The lines in the top-left, top-right, bottom-left, and bottom-right graphs correspond, respectively, to four values of the generalization parameter $b$, namely 0 (blue), 0.0577 (red), 0.115 (yellow), and 0.23 (magenta); to three values of the exploration parameter $\beta$, namely 10 (blue), 50 (red), and 500 (yellow); to three values of the discount rate $\gamma$, namely 0.3 (blue), 0.9 (red), and 0.99 (yellow); and to different numbers of allocation choices, namely 3 (blue), 6 (red), 11 (yellow), and 21 (magenta). There are 300,000 investors. For the remaining parameters, we set $L = T = 30$, $\alpha_{\pm}^{MF} = 0.3$, $\beta = 50$, $\gamma = 0.98$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.

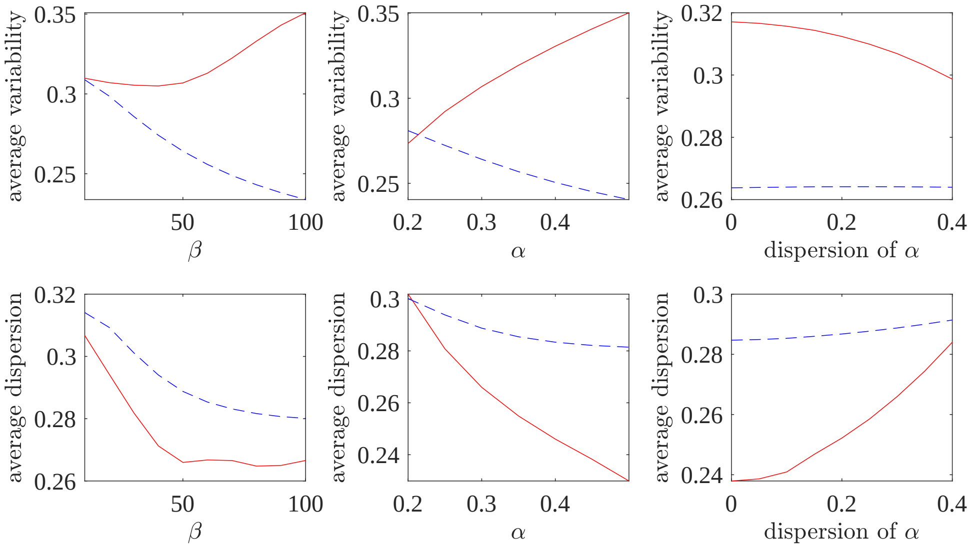


**Figure 4.** The upper graphs plot the variability of stock market allocations – the standard deviation of allocations between time 0 and time $T$, computed for each investor in turn and averaged across investors. The lower graphs plot the dispersion, across investors, of their stock market allocations at time $T$. The solid and dashed lines correspond to the model-based and model-free systems, respectively. For each system, the graphs vary the exploration parameter $\beta$, the mean learning rate $\bar{\alpha}$, and the dispersion in learning rates $\Delta$, while keeping the other parameter values fixed at benchmark levels. There are 10,000 investors. The benchmark parameter values are $L = T = 30$, $\bar{\alpha} = 0.3$, $\Delta = 0.3$, $\beta = 50$, $\gamma = 0.98$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

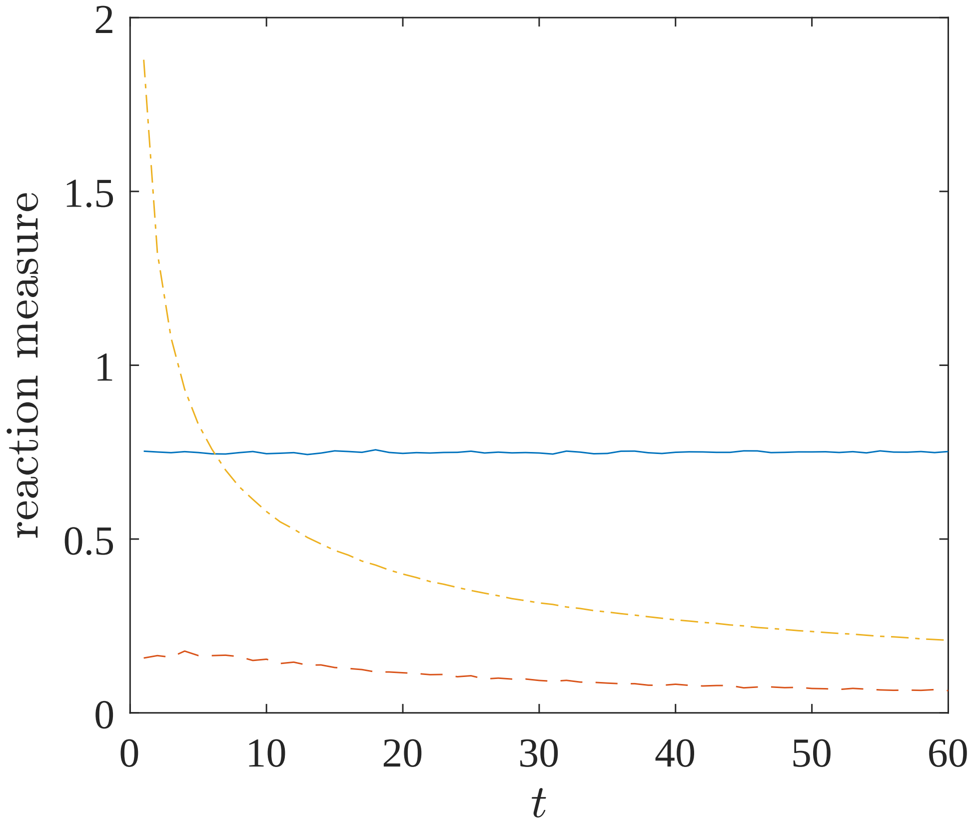


**Figure 5.** The graph plots, for each time from $t = 1$ to $t = 60$, the coefficient in a cross-sectional regression of the change in an investor's allocation at time $t$ on the stock market return at time $t$. The solid, dashed, and dash-dot lines correspond to the model-based system with a constant learning rate, the model-free system with a constant learning rate, and the model-based system with a declining learning rate, respectively. The dash-dot line is a rational benchmark: the area below it represents underreaction and the area above it, overreaction. There are 10,000 investors, each of whom is exposed to a different sequence of stock market returns. The parameter values are $L = 0$, $T = 60$, $\bar{\alpha} = 0.3$, $\Delta = 0.3$, $b = 0.0577$, $\gamma = 0.98$, $\mu = 0.01$, $\sigma = 0.2$, and $\beta = 50$ except in the case of the rational benchmark where there is no exploration.

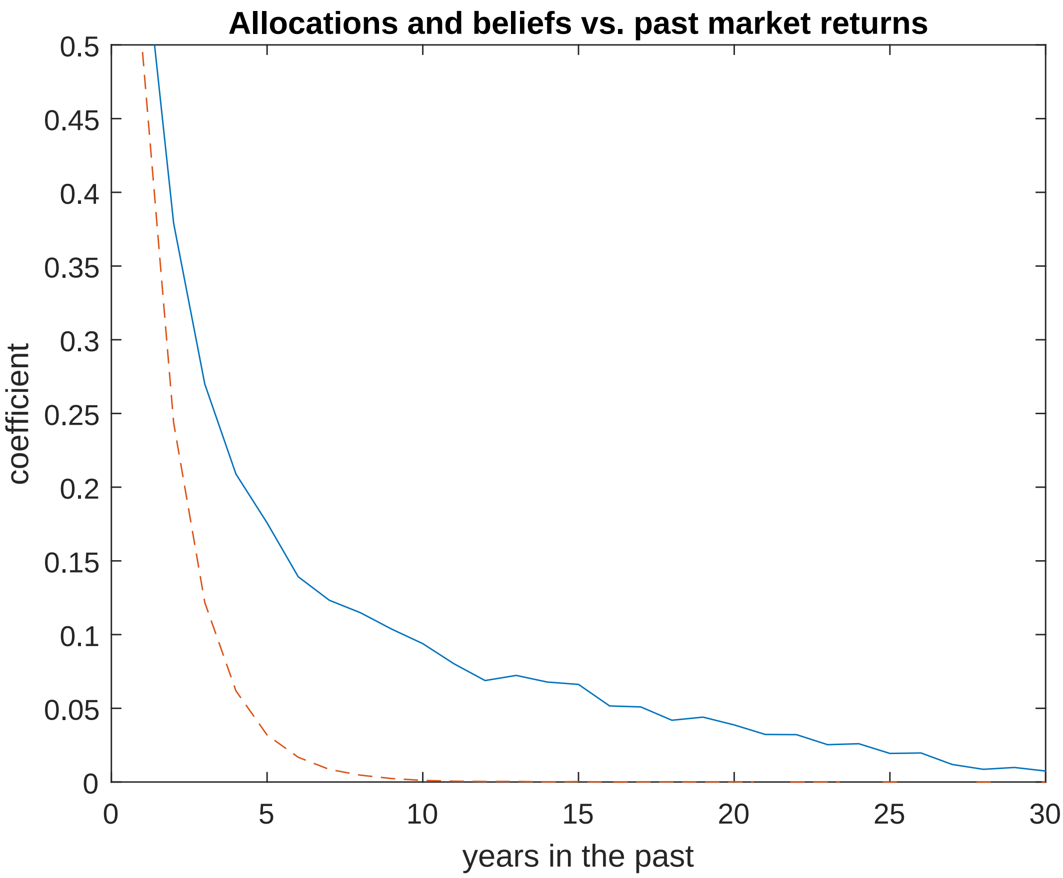


**Figure 6.** The solid line plots the coefficients in a regression of the stock market allocation $a_T$ at date $T$ chosen by investors who use a hybrid system to make decisions on the past 30 years of stock market returns the investors were exposed to. The dashed line plots the coefficients in a regression of investors' expectations at time $T$ about the future one-year stock market return on the past 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 $\bar{\alpha} = 0.5$ and width $\Delta = 0.3$. We also set $L = T = 30$, $\beta = 50$, $\gamma = 0.98$, $b = 0.0577$, $w = 0.5$, $\mu = 0.01$, and $\sigma = 0.2$.

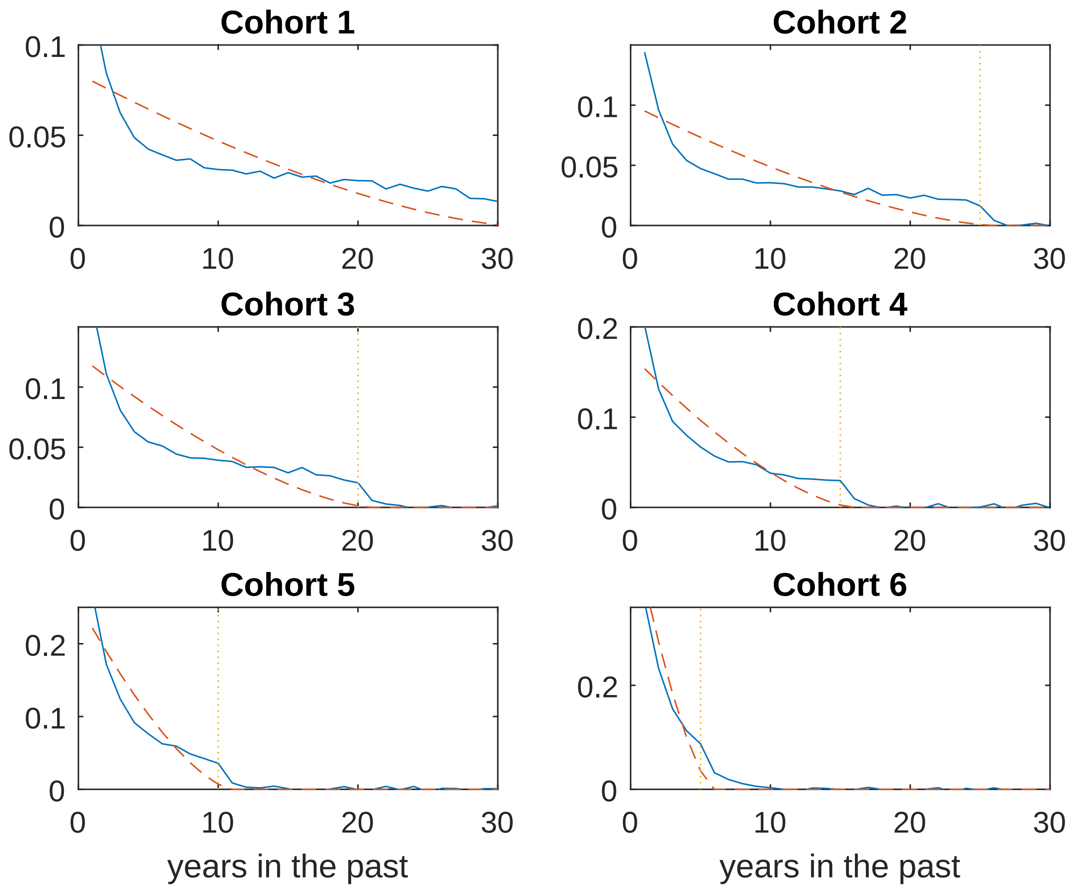


**Figure 7.** The six graphs correspond to six cohorts of investors. In each graph, the solid line plots the coefficients – normalized to sum to one – in a regression of the time $T$ stock market allocations $a_T$ of the investors in that cohort on the past returns they have been exposed to. The six cohorts have different numbers of years of experience, namely $n = 5$, 10, 15, 20, 25, and 30; the vertical dotted line in each graph marks the time at which the cohort enters financial markets. There are 300,000 investors, with 50,000 in each cohort. For each investor, each of $\alpha_+^{MF}$, $\alpha_-^{MF}$, $\alpha_+^{MB}$, and $\alpha_-^{MB}$ is drawn independently from a uniform distribution with mean $\bar{\alpha} = 0.5$ and width $\Delta = 0.3$. We also set $L = T = 30$, $\beta = 50$, $\gamma = 0.98$, $b = 0.0577$, $w = 0.5$, $\mu = 0.01$, and $\sigma = 0.2$. In each graph, the dashed line plots a functional form for experience effects proposed and calibrated to data by Malmendier and Nagel (2011), namely $(n-k)^\lambda/A$, where $k$ is the number of years in the past, $\lambda = 1.5$, and $A$ is a normalizing constant.