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

Nicholas Barberis and Lawrence Jin

August 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; and can help explain the dispersion in stock market allocations across investors as well as the inertia in these allocations over time. 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 values of different courses of action.

*This version is for conference submission only, not for broader distribution. 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 significantly 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 setting 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 host of facts in finance.[1]

The model-free and model-based learning algorithms operate in a specific dynamic setting. In this setting, at time t, the world is in state s_t and the individual takes an action a_t; at time $t+1$, he receives a stochastic reward R_{t+1} and arrives in state s_{t+1}. The individual's goal is to maximize the long-term sum of rewards.

The model-free and model-based systems both try to solve this problem by estimating an important 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 typically 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 probability distribution of s_{t+1} and R_{t+1} given s_t and a_t. Instead, it learns from the individual's experience: if, after the action a_t in state s_t, it observes the reward R_{t+1}, it updates its estimate of $Q(s_t, a_t)$, the value of taking action a_t in state s_t. To do the updating, it makes use 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 to receive; and a learning rate, which controls the extent of the updating. This model-free framework has been increasingly embraced 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 an estimate of $p(s_{t+1}, R_{t+1}|s_t, a_t)$, i.e., of the probability of a reward R_{t+1} and state s_{t+1} following action a_t in state s_t. A number of model-based approaches are possible, but we use one that is often adopted in research in psychology and that, like the model-free system, has neuroscientific support. In this

[1] An important early paper on this two-system 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).

framework, after observing reward R_{t+1} and state s_{t+1} following an action a_t in state s_t, the individual increases his estimate of $p(s_{t+1}, R_{t+1}|s_t, a_t)$ while downweighting the probabilities of other outcomes following s_t and a_t. To do the updating, the individual again uses a learning rate which controls the extent of the updating; and also a quantity called the state prediction error, which, analogous to the reward prediction error, is the difference between the probability with which (s_{t+1}, R_{t+1}) occurs, namely 1, and the prior probability the individual had assigned to it occurring. 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, humans use these two systems in combination (Glascher et al., 2010; Daw et al., 2011). Specifically, they take a weighted average of the $Q(s, a)$ values produced by each of the model-free and model-based systems, and then use the resulting "hybrid" $Q(s, a)$ to make a choice.

In this paper, we import this framework into a financial 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 is allocating money between a risk-free asset and a risky asset in order to maximize the expected log utility of wealth at some future horizon. We show that this problem fits the canonical context in which model-free and model-based learning algorithms are applied.

One difference between the model-free and model-based systems – a difference 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, the model-free system 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. To understand the behavior of the model-free and model-based systems, we start by studying them separately. In particular, we look at how the stock market allocation proposed by each of the model-free and model-based systems depends on the stock market returns that the individual has been exposed to. We find that, for both systems, the allocation puts weights on past returns that are positive and that decline for more distant past returns. However, one crucial property of the systems that we document – a property that plays a role in several applications – is that, for the model-free system, the weights that the stock market allocation puts on past market returns decline slowly as we go further into the past, so that the allocation puts substantial weight

even on distant past returns; by contrast, for the model-based system, the weights decline quickly, so that the allocation depends primarily on recent returns. We give the full intuition for these results in the main part of the paper.

We then use our framework to shed light on a range of empirical facts. Most fundamentally, the framework offers a foundation for extrapolative demand. Based on empirical evidence, several models propose that an individual's demand for an asset is a weighted average of past returns, with more weight on more recent returns (Barberis et al., 2015, 2018). In providing a foundation for such extrapolative demand, our framework delivers the new insight that this demand may have two distinct components operating at different frequencies – a model-free component which puts weight even on distant past returns, and a model-based component which puts weight primarily on recent 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 no weight on returns he has not experienced. Two features of our framework allow it to capture this. First, since the model-free system is in operation only when the individual is actively investing, the allocation proposed by the model-free system puts no weight at all on returns he has not directly experienced. Second, since the model-based system primarily puts weight on recent returns, it puts little weight on returns from the distant past, in other words, from before the time when the individual started investing. Taken together, these features imply that the allocation proposed by the hybrid system which combines the model-free and model-based systems puts substantially more weight on returns the individual has experienced than on returns he has not.

Our framework is also able to 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' actual 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.

In the main part of the paper, we discuss several other applications – for example, to explaining the low sensitivity of allocations to beliefs in the cross-section of investors (Giglio et al., 2021); and to understanding the cross-sectional dispersion in investor allocations to the stock market, as well as the individual-level inertia in these allocations over time.

In the standard implementation of model-free learning used by psychologists, the algorithm, at each time, updates only the value of the most recently taken action. However, there is evidence that the human model-free system can "generalize": after seeing the reward from taking an action a_t, it can use this experience to also update the values of other actions that are "similar" to a_t. We incorporate such generalization into our framework and show that it can improve the fit with the data. For example, it leads the model-free system to put even greater relative weight on distant past returns, thereby allowing us to explain all the better why investors' allocations do indeed put substantial weight on distant past returns.

Beyond the applications discussed above, our framework also points to some bigger themes. First, we find that the parameters of the framework that best fit the data put higher weight on the model-free system than on the model-based 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 which 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, 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 experience-weighted attraction learning – see Camerer and Ho (1999) – 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 had been written. Moreover, the financial setting we focus on is quite 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 incorporated into a simple portfolio-choice setting. 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 wide range of applications. Section 5 discusses some additional issues raised by our analysis, while Section 6 concludes.

2 Models

In the past decade, researchers in cognitive psychology and decision neuroscience have increasingly adopted a new framework for studying human behavior in experimental settings,

one that differs from that used by economists. This 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 – specifically, a simple portfolio-choice setting.

Model-free and model-based learning algorithms are intended to solve a specific decision problem, which we now describe. Time is discrete, and is 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 problem 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 $V_t(s_t)$ – 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 $V_t(s_t)$ 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 the probability distribution $p(\cdot)$; and even if we have a good sense of this distribution, it may be difficult to solve the Bellman equation for the value function $V_t(s_t)$.[4]

[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 $p(\cdot)$.

[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

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 situations, people do not appear to be able to intuit their way to the optimal solutions of economic problems, in that their behavior is not fully rational (Barberis and Thaler, 2003). Moreover, we have no understanding of why, in some cases, people do seem to be able to intuitively discern the optimal solution, while in other cases, they do not. As such, for any particular application, we cannot presume that people have managed to intuitively solve the Bellman equation. 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 two model-free algorithms most commonly used by psychologists to understand decision-making in experimental settings are known as Q-learning and SARSA. We focus on Q-learning here.[6]

In Section 2.1, we describe 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

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

model-based algorithms are combined. In Section 2.2, we present a simple portfolio-choice setting that fits the structure of the decision problem above. Much of the paper will be about exploring the properties and possible applications of model-free and model-based learning in this simple 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 $Q^*(s_t, a_t)$ be the value of taking action a_t in state s_t at time t and then continuing 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 $Q^*(s_t, a_t)$ accurately for all possible actions a_t, so that it can learn a good action to take in state s_t.

Suppose that, at time t in state s_t, the algorithm takes an action a_t – 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 $Q^*(s_t, a_t)$ is $Q^{\text{old}}(s_t, a_t)$. At time $t+1$, after observing the reward R_{t+1}, its estimate of $Q^*(s_t, a_t)$ is updated as follows:

$$Q^{\text{new}}(s_t, a_t) = Q^{\text{old}}(s_t, a_t) + \alpha^{MF}[R_{t+1} + \gamma \max_{a'} Q^{\text{old}}(s_{t+1}, a') - Q^{\text{old}}(s_t, a_t)], \qquad (3)$$

where α^{MF} is known as the learning rate – the superscript MF stands for model-free – and the term in square brackets is an important quantity known as the reward prediction error (RPE).

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

$$p(a_t = a) = \frac{\exp[\beta Q^{\text{old}}(s_t, a)]}{\sum_{a'} \exp[\beta Q^{\text{old}}(s_t, a')]}. \qquad (4)$$

Here, β is known as the "inverse temperature" parameter. 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. This probabilistic approach, often known as a "softmax" approach, serves an important purpose: it encourages the algorithm to "explore," in other words, to try actions other than the one that currently has the highest Q value in order to see whether

one of these other actions has an even higher Q value.

The algorithm is initialized at time 0 by setting $Q(s,a) = 0$ for all s and a. Consistent with (4), the time 0 action is chosen randomly from the set of possible actions. The process then proceeds as described in equations (3) and (4).

To see why equation (3) is a sensible updating rule, recall that the quantity $Q^*(s_t, a_t)$ satisfies the Bellman equation

$$Q^*(s_t, a_t) = E_t[R_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a')], \tag{5}$$

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 (3) as

$$Q^{\text{new}}(s_t, a_t) = (1 - \alpha^{MF})Q^{\text{old}}(s_t, a_t) + \alpha^{MF}[R_{t+1} + \gamma \max_{a'} Q^{\text{old}}(s_{t+1}, a')], \tag{6}$$

we see that the Q-learning algorithm is taking an estimate of the right-hand side of (5) and then updating $Q^{\text{old}}(s_t, a_t)$ in the direction of this estimate to an extent determined by the learning rate α^{MF}. Specifically, it proxies for the expected reward $E_t(R_{t+1})$ in (5) 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 optimal 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 that on the right-hand side of (3).[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

$$\begin{aligned} Q^{\text{new}}(s_t, a_t) &= Q^{\text{old}}(s_t, a_t) + \alpha_+^{MF}(\text{RPE}) \qquad \text{for RPE} \geq 0 \\ Q^{\text{new}}(s_t, a_t) &= Q^{\text{old}}(s_t, a_t) + \alpha_-^{MF}(\text{RPE}) \qquad \text{for RPE} < 0. \end{aligned} \tag{7}$$

[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).

In what follows, we also adopt this modification.

In the basic implementation of model-free learning described above, after taking an action a_t in state s_t, the algorithm updates only the Q value for that action-state pair. It is natural to ask whether the algorithm can "generalize" from its experience of (a_t, s_t) to also update the Q values of other action-state pairs. We return to this issue 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 learning algorithm described above. Before we do so, it will be helpful to first describe the financial decision-making problem that we will apply both algorithms to – a problem that fits the setting specified above.

We consider a simple portfolio choice problem, namely a choice between two assets: a risk-free asset and a risky asset which we think of as the overall 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

$$\begin{aligned} \log R_{m,t} &= \mu + \sigma\varepsilon_t \\ \varepsilon_t &\sim N(0,1), \text{ i.i.d.} \end{aligned} \tag{8}$$

At each time t, the investor chooses the fraction of his wealth that he allocates to the risky asset; this corresponds to the "action" in the framework of Section 2.1, so we use the notation a_t for it. The investor's goal is to maximize the expected log utility of his wealth at some future horizon. Because of liquidity needs whose timing is uncertain, he does not know in advance how far away this horizon is. At time 0, he enters financial markets. If, coming into time $t \geq 1$, the investor is still active 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 then to solve

$$\max_{\{a_t\}} E_0\left[\sum_{t=1}^{\infty} \gamma^{t-1} \log R_{p,t}\right], \tag{9}$$

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 (9), 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 (9). The solution – see the Appendix – is that, at each time t, the investor allocates a constant fraction a^* of his wealth to the stock market, where

$$a^* = \arg\max_a E_t \log((1-a)R_f + aR_{m,t+1}). \tag{10}$$

The fact that the problem in (9) can be solved mathematically does not necessarily mean that real-world investors will be able to quickly find their way to the solution in (10). As noted earlier, many investors may have a poor sense of the actual 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 (10) 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 – again, actions that will feel unnatural to many investors.

If investors are unable to explicitly compute the solution to the problem in (9), they may instead fall back 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-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 (6) directly. However, it is natural to start with the simplest case – the case with no state dependence, so that $Q(s_t, a_t)$ is replaced by $Q(a_t)$. We will find that even this simple case has rich implications that shed light on observed facts, so it will be our main focus.

What is the psychological justification for putting aside the state dependence? It is tempting to justify this assumption by saying that investors understand that asset returns are i.i.d.: in that case, the allocation problem has the same form at each time t, and dropping the dependence of $Q(s, a)$ on s is justified. The difficulty with this is that 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. As such, a better justification for replacing $Q(s,a)$ by $Q(a)$ is that it is a simplification on the part of the investor.

As in Section 2.1, then, let $Q^*(a)$ be the value of taking action a at some moment of time and then continuing optimally from the next period on. Suppose that, at time t, the individual takes the action 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

$$Q_{t+1}^{MF}(a) = Q_t^{MF}(a) + \alpha_{+/-}^{MF}[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)], \tag{11}$$

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

$$p(a_t = a) = \frac{\exp[\beta Q_t^{MF}(a)]}{\sum_{a'} \exp[\beta Q_t^{MF}(a')]}. \tag{12}$$

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

$$\begin{aligned} Q^*(a) &= \frac{1}{1-\gamma} E \log((1-a)R_f + aR_{m,t+1}) \quad \text{for } a = a^* \\ &= E \log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E \log((1-a^*)R_f + a^* R_{m,t+1}) \quad \text{for } a \neq a^*. \end{aligned} \tag{13}$$

In the basic model-free algorithm in equation (11), after taking action a_t at time t, we update only the value of the action taken. It is natural to ask whether the algorithm can "generalize" from its experience of taking the action a_t 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. We therefore incorporate generalization into our framework.

We have considered three types of generalization. All three approaches are motivated by research in computer science or psychology, and all three lead to similar results; we therefore focus on one of them here.

Given that we are working with the model-free system, it is important that the general-

ization we consider does not use any information about the structure of the environment or of the allocation problem. We therefore consider a simple form of generalization based on the notion of similarity: after choosing an allocation and observing the subsequent portfolio return, the algorithm updates not only the Q value of the chosen allocation, but also the Q values of similar allocations. We implement this as follows. After taking action a at time t, the algorithm updates the values of *all* allocations according to:

$$Q_{t+1}^{MF}(\widehat{a}) = Q_t^{MF}(\widehat{a}) + \alpha_{+/-}^{MF}\kappa(\widehat{a})[\log R_{p,t+1} + \gamma \max_{a'} Q_t^{MF}(a') - Q_t^{MF}(a)], \tag{14}$$

where

$$\kappa(\widehat{a}) = \exp(-\frac{(\widehat{a}-a)^2}{2b^2}). \tag{15}$$

In other 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 – and one that is all the lower, the more different $\widehat{a}$ is from a, to an extent determined by the Gaussian function in (15).

We will consider a range of values of b, but for our baseline analysis, we set $b = 0.0577$. This has a simple interpretation: for this b, the Gaussian distribution in (15) 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.

The applications we discuss in Section 4 do not hinge on generalization: they follow even for the basic model-free algorithm in (11) that features no generalization at all. We incorporate generalization because it is psychologically reasonable and because it will turn out to have rich implications of its own, some of which will improve the 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 (8): by its model-free nature, it does not have a model of the environment. Indeed, from the perspective of the algorithm, it has no idea what a "risk-free asset" is or what the "stock market" is. It is simply choosing different actions – different combinations of these unfamiliar objects – seeing what reward they deliver, and then updating the values of the chosen actions.

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 make use of information about the probability distribution of future rewards conditional on past actions. There are various model-based frameworks we can consider. One possibility is to take as the model-based framework the optimal Q^* in (13). Here, we instead use another framework that, like the model-free framework of Section 2.1, is based directly on a framework that is used extensively by psychologists and is supported by neural evidence from decision-making experiments.

In the framework we consider, the 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 that are analogous to the reward prediction errors of Section 2.1 and are sometimes referred to as "state prediction errors." Specifically, suppose that the investor observes a stock market return $R_{m,t} = R$ at time t, and suppose also that, before observing the return, the prior probability he assigned to it occurring was $p^{\text{old}}(R_{m,t} = R)$. At time t, he then updates the probability of this return as:

$$p^{\text{new}}(R_{m,t} = R) = p^{\text{old}}(R_{m,t} = R) + \alpha^{MB}[1 - p^{\text{old}}(R_{m,t} = R)], \tag{16}$$

where α^{MB} is the learning rate for the model-based framework. The term $1-p^{\text{old}}(R_{m,t} = R)$ can be thought of as a prediction error: the investor's prior estimate of the probability of the return equaling R was $p^{\text{old}}(R_{m,t} = R)$; 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 $1 - \alpha^{MB}$ 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, $p^{\text{old}}(R_{m,t} = R) = 0$, which simplifies (16) to

$$p^{\text{new}}(R_{m,t} = R) = \alpha^{MB}.$$

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 this notation, a comma separates a return realization from its perceived probability, while semicolons separate the different return outcomes:

$$(R_1, 1)$$

$$
\begin{aligned}
&(R_1, 1-\alpha^{MB}; R_2, \alpha^{MB}) \\
&(R_1, (1-\alpha^{MB})^2; R_2, \alpha^{MB}(1-\alpha^{MB}); R_3, \alpha^{MB}) \\
&(R_1, (1-\alpha^{MB})^3; R_2, \alpha^{MB}(1-\alpha^{MB})^2; R_3, \alpha^{MB}(1-\alpha^{MB}); R_4, \alpha^{MB}). \qquad (17)
\end{aligned}
$$

The above approach is motivated by papers in decision neuroscience that adopt a similar model-based framework; these include 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 that in square brackets in (16) (Glascher et al., 2010).

We noted in Section 2.1 that, when they implement model-free learning, psychologists allow for different model-free learning rates, α_+^{MF} and α_-^{MF} for positive and negative reward prediction errors, respectively. We extend the above model-based algorithm in a similar way, allowing for different model-based learning rates, α_+^{MF} and α_-^{MF}, depending on whether the latest net stock market return is positive or negative, so that

$$
p^{\text{new}}(R_t = R) = \alpha_+^{MB} \text{ for } R \geq 1, \qquad (18)
$$

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

$$
p^{\text{new}}(R_t = R) = \alpha_-^{MB} \text{ for } R < 1, \qquad (19)
$$

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

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 the equations in (13) for the correct value of $Q^*(a)$, and applying them for his *perceived* time t return distribution:

$$
\begin{aligned}
Q_t^{MB}(a) &= \frac{1}{1-\gamma} E_t^p \log((1-a)R_f + aR_{m,t+1}) \quad \text{for } a = a^* \qquad (20) \\
&= E_t^p \log((1-a)R_f + aR_{m,t+1}) + \frac{\gamma}{1-\gamma} E_t^p \log((1-a^*)R_f + a^* R_{m,t+1}) \quad \text{for } a \neq a^*,
\end{aligned}
$$

where

$$
a^* = \arg\max_a E_t^p \log((1-a)R_f + aR_{m,t+1}) \qquad (21)
$$

and where (20) differs from (13) 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.

2.4 A hybrid model

Current research 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, the individual computes a "hybrid" value of $Q(a)$ that is a weighted average of the model-free and model-based Q values:

$$Q_t^{HYB}(a) = (1-w)Q_t^{MF}(a) + wQ_t^{MB}(a), \tag{22}$$

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

$$p(a_t = a) = \frac{\exp[\beta Q_t^{HYB}(a)]}{\sum_{a'} \exp[\beta Q_t^{HYB}(a')]}. \tag{23}$$

The model-free and model-based systems differ most fundamentally in the way 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 the two systems that will play a role in one of our applications: 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 investing 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 above in (18) and (19); and computes Q^{MB} values based on that distribution, as in (20).[8]

[8]Having the model-free and model-based algorithms apply over different time intervals is not required for most of our applications in Section 4 – but it is crucial for one application, discussed in Section 4.2. Our implementation here is consistent with evidence in decision neuroscience. For example, Dunne et al. (2016) conduct an experiment in which participants actively experience slot machines that deliver a stochastic reward, but also passively observe other people playing the slot machines. fMRI measurements show that,

3 Model Properties and an Example

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

To conduct the analyses, 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 operate only from time 0 to time T. 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 from $t = -L$ to $t = T$. We think of each time period as one year and set $L = T = 30$. Before he starts investing at time 0, then, an investor has access to 30 years of prior data going back to $t = -30$. We then track his allocation decisions over the next 30 years.[9]

The four learning rates, α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB}, play an important role in our framework. How should they be set? Since we are taking a "positive" perspective – our goal in this paper is to explain observed behavior – what matters for our purposes is not the learning rates people should choose, but rather the learning rates they actually choose. In practice, learning rates are likely to depend on attention – greater attention corresponds to a higher learning rate – and on an individual's perception of the stability of the environment: if he perceives the environment to be changing over time, he will prefer a constant learning rate that puts more weight on recent observations. Psychology research does not offer definitive guidance on the learning rates people use, but several studies find evidence for learning rates that are relatively constant over time, even when these rates should optimally be time-varying. For this reason, and because this is the simplest assumption we can make, we focus on learning rates that are constant over time. To start, we give all investors the same constant learning rate. Later, we allow for dispersion in these rates across investors.

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."

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

3.1 The mechanics of the two systems

To show how the model-free and model-based systems work, we start with an example. We consider a single 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 (8) with $\mu = 0.01$ and $\sigma = 0.2$, which provides a reasonable fit to historical annual U.S. stock market returns. We set the investor's learning rates to $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, and also set $\beta = 50$, $\gamma = 0.9$, and $b = 0.0577$. At each time, we allow the investor to choose his stock market allocation a_t from one of 11 possible allocations $\{0\%, 10\%, \ldots, 90\%, 100\%\}$. Later in the paper, we will examine how the coarseness of the action set affects our 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: 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 (12), (14), and (15) (upper panel) and the model-based Q values, Q^{MB}, based on equation (20) (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 bold number -0.051 at date 1 in the upper table means that the investor chose an allocation of 90% at date 0.[10]

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 90%. The net stock market return at time 1 is negative, which means that the investor's portfolio return and reward prediction error are also negative. The time-1 Q value for the 90% allocation therefore falls below zero. As per equation (15), the algorithm also engages in some generalization: since an 80% allocation and a 100% allocation are similar to a 90% 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.

[10]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 (12). 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.

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 chooses the allocation 50%. Since the market falls slightly at time 3, the net portfolio return at that time is slightly negative, but since the reward prediction error as a whole is slightly positive, the time-3 Q value of the 50% allocation goes up by a small amount. At dates 3 and 4, the investor chooses allocations of 20% and 0%, respectively, and updates the values of these allocations and their close neighbors based on the prediction errors they lead to at dates 4 and 5.

The lower panel shows that the Q values generated by the model-based system are quite different. By time 0, the model-based system has already been operating for 30 periods, and so already has well-developed Q values for each of the 11 allocation strategies. In the periods immediately preceding time 0, the simulated stock market returns are slightly 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 is what makes the model-free system powerful and an 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, it can be inefficient.

We noted earlier that, under some conditions, the model-free and model-based algorithms in Sections 2.1 and 2.3 both converge to the optimal allocation in (21). Why, then, do the investors in our framework not follow this policy as soon as they start investing at time 0? The main reason is that, even with optimal learning rates, convergence takes a long time – far longer than even the 30-year time interval over which we track investors. In addition, convergence usually requires learning rates that decline over time, as opposed to the constant learning rates that we use here.

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. To understand this, we take 50,000 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 $\{R_{m,T}$, $R_{m,T-1},\ldots,R_{m,T-29}\}$ investors have been exposed to, and record the coefficients.

We conduct this analysis separately for four cases: one where investors' allocations are determined only by the model-free system; one where they are determined only by the model-based system; one where they are determined by the hybrid system; and one where they are determined by a "pure" model-free system without generalization, which updates Q values according to equation (11) instead of (14). In each case, we again take $L = 30$ and $T = 30$. All investors have the same constant learning rates $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, although we will consider other learning rates as well. As before, we take $\beta = 50$, $\gamma = 0.9$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$ except in the case of the pure model-free system where there is no generalization. Finally, we set $w = 0.5$, so that the hybrid system puts equal weight on the model-free and model-based systems.

FIGURE 1 presents the results. The solid line marked "model-free" plots the coefficients on past returns in the regression described above for the case where allocations are determined by the model-free system; the left part of the graph corresponds to more recent returns. The two other solid lines plot the coefficients for the model-based and hybrid systems. Finally, the dashed line plots the coefficients for the "pure" model-free system.

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. Given that the hybrid system combines the model-free and model-based systems, it is natural that, as shown in the figure, the allocations under the hybrid system also depend positively on past returns, with declining weights.

FIGURE 1 also illustrates another important property of the model-free and model-based systems, one that plays a key role in some of our later applications. While both the model-free and model-based allocations put weights on past returns that are positive and decline the further back we go into the past, the decline is much faster in the case of the model-based system. Put differently, while the model-based system puts heavy weight on recent returns,

the model-free system puts substantial weight even on distant past returns.

We now explain 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 in the future.

The intuition for the model-free system is more subtle. We present it for the case of pure model-free learning and discuss generalization later in the section. 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. However, the reinforcement 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. Overall, then, a good stock market return leads the investor to choose a higher stock market allocation, on average. 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.

Why do the coefficients decline more slowly for the model-free system as we go further into the past? We again give the intuition for pure model-free learning and discuss generalization below. The model-free system learns slowly: after taking an action a_t at time t, it updates only the value of action a_t; put differently, at time t, the most recent stock market return is reflected only in the Q value of action a_t. As a result, the full set of 11 time-t Q values for the 11 allocations do not depend substantially more on recent returns as compared to distant returns – and so neither does the allocation chosen. By contrast, the model-based system learns quickly: after observing the most recent stock market return at time t, it updates the Q values of all 11 allocations. These 11 Q values therefore put substantial weight on the most recent return as compared to more distant returns, and therefore the investor's allocation does too.

These properties are robust to changing the value of the learning rates. The top (bottom) graph in FIGURE 2 shows how the allocation recommended by the model-based (model-free) system depends on past stock market returns for four different values of the learning rates,

namely 0.05, 0.1, 0.2, and 0.5. The graphs show that, for both systems, regardless of the learning rate, the allocations put weights on past returns that are positive and declining the further back we go into the past. They also show that, as we raise the learning rate, both systems put more weight on more recent returns.

FIGURE 1 shows that, for $b = 0.0577$, generalization has only a small impact on the relationship between the model-free allocation and past returns: the dashed and solid model-free lines are similar. However, if we look at a wider range of values for b, we obtain a richer set of results. 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 b: 0.001, 0.0577, 0.115, and 0.23. The first of these values approximates the pure model-free system without generalization; the other three values give the Gaussian distribution in (15) 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 allocations and past returns for three different values of β, which controls the degree of exploration, namely 10, 50, and 500. Recall that, as β rises, the investor explores less; in other words, he is more likely to choose the allocation with the highest estimated Q value. To understand the effect of β more clearly, we switch off generalization in this graph by setting b to 0.001.

The graph shows that, for a wide range of values of β, the relationship between the allocation and past returns is qualitatively similar to that for our benchmark case of $\beta = 50$. However, when β is very high – based on our computations, higher than 200 – the weights on past returns *increase* for more distant 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 the initial allocation chosen. If the value of β is high, the investor is likely to stick with this allocation for a substantial period of time. As such, the early returns he experiences have an outsize effect on his subsequent allocations.

The bottom-left graph also plots the relationship between allocation and past returns for the same three values of β, but now re-introduces generalization at our baseline level of $b = 0.0577$. The graph shows patterns similar to those in the top-right graph. The presence of generalization means that somewhat lower values of β – values in excess of – are enough to generate an increasing relationship between allocations and returns.

The bottom-right graph plots the relationship between allocations and past returns for three different values of the discount rate γ, namely 0.3, 0.9, and 0.99. We obtain qualitatively similar patterns for a range of values of γ, but the decline in weights is steeper for lower levels of γ.

FIGURE 3 shows that a greater degree of generalization and a lower level of exploration can create a relationship between allocations and past returns that puts greater weight on distant past returns. There is little evidence of such "primacy" effects in investors' actual allocations, which suggests that the level of generalization in the model-free system is lower than that produced by $b = 0.23$, and that the inverse temperature parameter β is not especially high. Nonetheless, a lesson of FIGURE 3 is that, if a person does exhibit a primacy effect, this may be due to his model-free system engaging in more generalization, or in less exploration.

In the simulations we conducted above, all $50,000$ investors have the same learning rates. When we turn to applications in the next section, we will allow for dispersion in learning rates across investors. For now, we note that allowing for such dispersion does not have much impact on the relationship between allocations and past returns. For example, if each investor draws each of his learning rates independently from a uniform distribution centered at 0.3, such as $[0.1, 0.5]$, the relationship between allocations and past returns is similar to that shown in FIGURE 1 for the case where all investors have the same learning rate of 0.3. We present quantitative results on this issue in the next section.

Thus far, we have looked at how investors' allocations depend on past stock market returns. We have also examined how their allocations depend on their past *portfolio* returns and their past allocations. For each of the model-based and model-free systems in turn, we take the stock market allocations recommended by the system and regress them on investors' past experienced portfolio returns $\{R_{p,T-j}\}_{j=0}^{29}$; separately, we also regress them on investors' past allocations $\{a_{T-j}\}_{j=1}^{30}$. We find that, in the case of the model-based system, it is past stock market returns that have the greatest explanatory power for final allocations, followed by experienced portfolio returns and past allocations. In the case of the model-free system, it is the opposite: the past allocation has the greatest explanatory power, followed by experienced returns, followed by stock market returns.

3.3 Dispersion and variability in allocations

In this section, we consider some other properties of the model-free and model-based systems – properties related to the dispersion and the variability in investor allocations – that will play a role in the applications we discuss in Section 4. 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 find – in broad terms; we put concrete numbers to this below – that the model-free system generates more dispersion than the model-based system, but less variability. Moreover, these findings are connected: the model-free system generates more dispersion in part *because* it produces less variability.

Consider a setting with $50,000$ investors, all of whom have the same learning rates. As before, $L = T = 30$, $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, $\gamma = 0.9$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$. For now, suppose that the inverse temperature parameter β is high – equal to 200, say. We expose all investors 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 many times for different return sequences and average the resulting set of dispersion and variability estimates. We find that the model-based system generates a low dispersion of 11.4%, while the model-free system generates a much higher dispersion of 27.8%.

The model-based system generates low dispersion because, at each date, all investors update their perceived return distributions in the same way and compute the same Q values for all 11 allocations; hence, aside from a small amount of experimentation, they choose the same allocation in each period.

The model-free system, by contrast, generates much higher dispersion. This stems from the combination of the probabilistic action choice and the reinforcement inherent in the model-free system. At time 0, the probabilistic action choice in (12) 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 across investors.

While the model-free system generates higher dispersion than the model-based system, it generates lower variability – specifically, variability of 17.5%, as compared to 39.9% for the model-based system. In each period, the model-based system leads investors to increase their allocation following a good return and to 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 a particular allocation to the stock market and the market then performs well, the Q value of the allocation will be pushed up, raising the chance that the investor will keep choosing this allocation in subsequent years.

FIGURE 4 illustrates these points. The top (bottom) graph plots the allocation chosen by the model-free (model-based) system for 10 randomly-chosen investors over the 31 dates from $t = 0$ to $t = T = 30$. The top graph shows the low variability and high dispersion generated by the model-free system, while the lower graph shows the high variability and low dispersion generated by the model-based system.

In the above analysis, we used a high value of β to illustrate the results more clearly. If we lower β toward our baseline level of 50, the patterns described above continue to hold, but to a lesser extent: the model-free system continues to generate lower variability and higher dispersion than the model-based system, but in a quantitatively less dramatic way.

In the above analysis, all investors have the same learning rates. If there is instead dispersion across investors in the learning rates, this will naturally lead to greater dispersion in allocations at time T. We return to this in Section 4.

4 Applications

In this section, we 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 explain a good deal about investor behavior in these markets.

We start by showing that a simple parameterization of the framework can qualitatively, and in some cases quantitatively, address a range of facts about investor behavior. By "simple," we mean that, in this parameterization, each investor's learning rates α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} are constant over time; and, for all investors, the values of these learning rates are drawn from the same distribution. 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. At the end of the section, we build on this analysis to propose model parameter values that can 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 two modifications whose purpose is to make the framework more realistic. First, we allow for dispersion across investors in the values of their learning rates. Second, we allow for different cohorts of investors who enter financial markets at different times. Specifically, we take $L = T = 30$ and consider six cohorts, each of which contains $50,000$ investors, 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 from time $t = -L$ to time $t = T$, but their model-free systems operate only from time $t = 0$ to time $t = T$. 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 from $t = -L$ to $t = T$, but the model-free system operates only from time $t = 5$ to $t = T$. In the same way, the four remaining cohorts enter at dates $t = 10$, 15, 20, and 25; we track their decisions until time $t = T$. For the sixth cohort, for example, the model-based system operates over the full timeline from $t = -L$ to $t = T$, but the model-free system only from $t = 25$ to $t = T$.

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 (9) and chooses allocations according to the hybrid system in (22)-(23). 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 α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} independently from the uniform distribution $[0.1, 0.5]$. We set $\beta = 50$, $\gamma = 0.9$, and $b = 0.0577$. Finally, we set $w = 0.2$, so that investors put more weight on the model-free system. Indeed, one of our striking findings is that the framework fits the data better when investors put more weight on the less informed system.

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 nonetheless 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, in other words, that their demand for a financial asset depends positively on the asset's past returns, and especially on its recent past returns. A common specification posits that asset demand is a weighted average of past returns where the weights are positive and lower for more distant past returns (Barberis et al. 2015, 2018).

The framework of Section 2 provides a foundation for such extrapolative demand. To confirm this, we run a regression of investors' allocation a_T at time T, as determined by the hybrid system, on the past stock market returns each of them has observed. The relationship between the allocation and past returns is plotted as the solid line in FIGURE 5. The plot 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 5 is similar to the line marked "Hybrid" in FIGURE 1 in that both lines correspond to decisions made under the Hybrid system. However, the two lines differ in that, relative to the analysis of Section 3, we are now allowing for dispersion across investors in their learning rates and for multiple cohorts. The multiple cohorts in particular make the solid line in FIGURE 5 decline more quickly than the "Hybrid" line in FIGURE 1: some of the investors 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 a new 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 primarily puts weight on *recent* returns; and one that stems from a model-free system that puts substantial weight even on *distant* past returns. This follows from the analysis in Section 3. As we showed there, if an investor has learning rates α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} that are constant over time and equal to one another, this leads both the model-free and model-based systems to recommend actions that depend positively on past returns, with weights on these returns that decline the further back we go into the past; importantly, however, the weights decline more quickly in the case of the model-based system.

If we raise the mean of the distribution from which each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} is drawn, then the dependence of an investor's allocation on past returns will be qualitatively similar to that shown by the solid line in FIGURE 5, but will put more weight on more recent past returns: as shown in FIGURE 2, higher learning rates lead both the model-free and model-based systems to put more weight on recent returns.

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. However, it is not "automatic" that the framework will generate an experience effect – as we explain below, it depends on the model parameters.

There are two key features of experience effects that we are hoping 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

$$(n-k)^{\lambda}/\Delta, \tag{24}$$

where λ is estimated to be approximately 1.5 and Δ is a normalizing constant, 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 5 presents the results. The six graphs correspond to the six cohorts. In each graph, the solid line plots the coefficients in the above regression; the dashed line plots the functional form in (24) 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 line and the dashed line, we see that our framework can capture both aspects of experience effects. Consider the bottom-right graph for cohort 6, which enters at time 25. 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 five years in the past – to one they did not experience, the return six years in the past. This jump has two sources. First, these investors' model-free systems put zero weight on returns more than five years in the past. Second, these investors' model-based systems put weight primarily on recent returns, and hence little weight on returns more than five years in the past. The little weight that, according to the solid line in the graph, these investors do put on returns more than five years in the past comes from the model-based system. The graph also shows that, within the subset of returns that these investors experience – the past five years of returns – their allocation puts greater weight on more recent past returns. Both the model-free and model-based systems contribute to this pattern.

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, they put more weight on more recent returns.

What kinds of parameter values lead to stronger experience effects, where by "stronger" we mean a larger jump in the weights that an investor's allocation puts on past returns as we move from a return he has experienced to one that he has not. We obtain stronger experience effects for *lower* model-free learning rates α_+^{MF} and α_-^{MF}. If these learning rates are high, then the model-free allocation will put heavy weight on recent returns; as such, if the investor entered financial markets at time $t << T$, the model-free allocation will put close to zero weight on both $R_{m,t}$ and $R_{m,t+1}$, so that there is no jump in the weights as we move from a return the investor experienced to one he did not. By contrast, we obtain stronger experience effects for *higher* model-based learning rates. If these rates are low, then the model-based allocation will put weight even on distant past returns: if the investor entered the markets at time $t << T$, it will put substantial weight on both $R_{m,t}$ and $R_{m,t+1}$, again

reducing any jump in weights as we move from a return the investor experienced to one he did not. Finally we obtain stronger experience effects for lower weights w on the model-based system: after all, it is the model-free system that helps to drive a wedge between experienced and non-experienced returns. As noted above, then, it is not automatic that our framework will generate experience effects; it depends on the values of key model parameters.

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 primarily on *recent* past returns – see Greenwood and Shleifer (2014) – studies of investor stock market allocations find that these depend even on distant past returns (Malmendier and Nagel, 2011).

Two fundamental features of the framework of Section 2 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 depend on relatively 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 4 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.

What kinds of parameter values produce a sharper frequency disconnect, in other words, lead to allocations that put substantially more weight on distant past returns than do investor beliefs? This will occur for *lower* model-free learning rates α_+^{MF} and α_-^{MF}: low values of these parameters lead the model-free system to recommend allocations that depend on distant past returns. It will be also be the case for *higher* model-based learning rates α_+^{MB} and α_-^{MB}: high values of these parameters lead investor beliefs to depend primarily on recent past returns. Finally, it will be the case when there is a low weight w on the model-based system, as this will lead allocations to be based more on the model-free system and hence to depend more on distant past returns. Overall, the frequency disconnect points in the same direction as experience effects: to low values of α_+^{MF} and α_-^{MF}, to high values of α_+^{MB} and α_-^{MB}, and to a low value of w.

We can compare the quantitative results in FIGURE 4 to those reported in empirical research on investor beliefs. Greenwood and Shleifer (2014) run a regression of expected one-year stock market returns on past one-year stock market returns and obtain a coefficient of approximately 0.1. FIGURE 4 shows that, for our parameterization, we obtain a coefficient of 0.3, which is higher than the empirical estimate. Meanwhile, FIGURE 4 shows that the coefficients on past returns decline by a multiplicative factor of approximately 0.7 for each year that we go into the past. Barberis et al. (2015) estimate a roughly similar rate of decline of 0.6 in the data, while Greenwood and Shleifer (2014) estimate a faster rate of decline, namely 0.6 on a quarterly basis.

4.4 Investor beliefs and the cross-sectional disconnect

Giglio et al. (2021) document another disconnect between investor beliefs and investor actions. They regress investors' allocations to the stock market on investors' expected one-year stock market returns. They obtain a coefficient approximately equal to one; the precise value depends on the specification. However, a traditional Merton model of portfolio choice suggests that, for reasonable levels of risk aversion, the coefficient should be substantially higher.

The framework of Section 2 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 the investor's 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. An investor's expectation about the future stock market return will then go up significantly: the model-based system, which determines beliefs, puts substantial

weight even on recent past 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 hence less weight on the most recent return.

We now examine this quantitatively. The fourth column in TABLE 2 presents, for a number of different parameterizations, the coefficient in a regression of the investor's allocation a_T to the stock market at time T on his expected return on the stock market over the next year. The first row of the table corresponds to the benchmark parameterization we have been using, namely one where the mean learning rates across investors equal 0.3 and the weight on the model-based system is 0.2. The other rows consider alternative mean learning rates and alternative weights w.

The table shows that, for the benchmark parameter values we have been using, the regression coefficient, namely 0.98, is similar to that obtained by Giglio et al. (2021). It also shows that the sensitivity of the allocation to beliefs falls when investors put a lower weight w on the model-based system, and when the mean model-based learning rates are higher. The sensitivity does not depend to any significant extent on the mean model-free learning rates. As such, the cross-sectional disconnect has implications for the parameter values that are broadly similar to those of experience effects and the frequency disconnect: it points to higher model-based learning rates and a lower weight w on the model-based system.

What is the appropriate benchmark to compare the numbers in the fourth column of TABLE 2 to? One benchmark is the fully optimal allocation in (10). While (10) does not allow us to derive an analytical expression for the sensitivity of allocation to beliefs, we can compute this sensitivity numerically. If we change the mean stock return by changing μ in (8) from -0.005 to 0 to 0.005 to 0.01, we obtain stock market allocations of 37%, 49%, 64%, and 74%, respectively. These numbers indicate a sensitivity of approximately 25, much higher than the sensitivity in the fourth column of TABLE 2, and generating a strong disconnect between allocations and beliefs, as in Giglio et al. (2021).

However, the solution in (10) may not be the ideal benchmark because it assumes that the investor has full knowledge of the return generating process. A better benchmark may be the model-based system itself. Accordingly, we take the allocations that would be recommended by the model-based system alone, and compute their sensitivity to investor beliefs. The results are in the fifth column of TABLE 2. We see that the sensitivities in the fourth column for the allocations based on the hybrid system are lower than the sensitivities in the fifth column, consistent with the disconnect identified in Giglio et al. (2021).

4.5 Dispersion and inertia in household allocations

Households differ a lot in their asset allocation decisions: some households participate in the stock market, while others do not; and among households that participate, the fraction of wealth invested in the stock market varies substantially. It is not easy to explain these differing allocations: regressions of allocations on various explanatory variables have a relatively 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 α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{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 at time T, a_T, on the differences in investors' 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 this to be a substantial 15%.

The second possibility is one discussed earlier in the paper in connection with FIGURE 4. The upper 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' observed 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. However, the framework in this paper offers a new way of thinking about it: it says that the inertia arises endogenously from the model-free system.

As noted in Section 3.2, the model-free system generates inertia for the following reason: 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. The upper panel in FIGURE 4 illustrates this inertia for three randomly-chosen investors: it shows that even though these investors are free to change their allocations at each date, they nonetheless each choose to

stick with the same allocation for several years in a row.

4.6 Individual and time fixed effects in investor beliefs

Giglio et al. (2021) find that there is wide variation, across investors and over time, in beliefs about the future stock market return. They examine the extent to which this variation can be captured by individual fixed effects and by time fixed effects. They find that individual fixed effects capture a larger share of the variation than time fixed effects.

To examine this in our setting, we take the beliefs of all investors in the last five periods of our economy, in other words, at dates $t = 26$, 27, 28, 29, and 30. Note that the investor composition remains the same across all five dates: all six cohorts are in the market for these five dates. With these five periods of beliefs, we run the same regression analyses as Giglio et al. (2021): one with individual fixed effects alone; one with time fixed effects alone; and one with both individual and time fixed effects:

$$\begin{aligned} EXP_{i,t} &= \phi_i + \varepsilon_{i,t} \\ EXP_{i,t} &= \psi_t + \varepsilon_{i,t} \\ EXP_{i,t} &= \phi_i + \psi_t + \varepsilon_{i,t}, \end{aligned} \tag{25}$$

where $EXP_{i,t}$ is investor i's expectation of the future one-year stock market return at time t, ϕ_i represents individual fixed effects, and ψ_t represents time fixed effects. To match Giglio et al. (2021), we run the analysis slightly differently than in earlier sections. We take $3,000$ investors, with 500 in each cohort; expose all $3,000$ to the same sequence of stock market returns from $t = -L$ to $t = T$; and then run the regressions in (25), recording the R^2 of each one. We then repeat this experiment many times, each time exposing the $3,000$ investors to a new sequence of stock market returns. In the results below, we report the average R^2 across these different simulations.

TABLE 3 presents the R^2 for the above three regressions for various values of the mean model-based learning rates $\overline{\alpha}_+^{MB}$ and $\overline{\alpha}_-^{MB}$, and of the dispersion of learning rates across investors. Since this analysis pertains purely to beliefs, the results depend only on the model-based learning rates. The table shows that there are two ways in which, in our framework, individual fixed effects can explain a large fraction of the variation in investor beliefs. The first is if the mean model-based learning rates $\overline{\alpha}_+^{MB}$ and $\overline{\alpha}_-^{MB}$ are low (the first and second rows of the table). If these rates are instead high, investor beliefs will react a lot to the most recent stock market return; time fixed effects will then capture much of the

variation in investor beliefs. The second possibility, illustrated by comparing the third and fourth rows of the table, is that there is wide dispersion across investors, in the learning rates.

The evidence on experience effects, the frequency disconnect, and the cross-sectional disconnect discussed above in Sections 4.2, 4.3, and 4.4 all point to model-based learning rates that are relatively high. This suggests that the second of the two possibilities is the more likely explanation, within our framework, for why individual fixed effects explain a large fraction of the variation in investor beliefs: there must be substantial dispersion, across investors, in their model-based learning rates.

The table shows that, even if we allow for substantial dispersion in model-based learning rates, our framework still does not fully explain the fraction of variation captured by individual fixed effects. This is not surprising – there are likely several factors, not captured by our framework, that affect a person's average optimism or pessimism about the stock market.

4.7 Inferring beliefs from the model-free system

Until now, we have associated beliefs purely with the model-based system. However, it is possible that investors also use the model-*free* system to make some inferences about beliefs. When an investor is surveyed for his beliefs – for example, his beliefs about the future stock market return or future risk of the stock market – 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 will naturally first consult their model-based system, which will give 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.

4.8 Parameter values

In the preceding sections, we took a simple parameterization – one where the mean model-free learning rate across investors equals the mean model-based learning rate – and showed that it can qualitatively address a range of facts about investor behavior, and, in some cases, can offer a reasonable quantitative fit to the observed data. Through our analysis of these applications, we have obtained a clearer sense of the parameter values that will offer an even better quantitative match to the data.

Specifically, the empirical facts on the relationship between past stock market returns and investor expectations of future returns – for example, the evidence in Greenwood and Shleifer (2014) – point to model-based learning rates α_+^{MB} and α_-^{MB} in the range from 0.1 to 0.3. The evidence of Giglio et al. (2021) discussed in Section 4.6, namely, that individual fixed effects explain much of the variation in investor beliefs, indicates that there is wide dispersion across investors, in the values of their model-based learning rates α_+^{MB} and α_-^{MB}. Finally, the evidence on experience effects and the disconnect between allocations and beliefs in the frequency domain and in the cross-section – the evidence discussed in Sections 4.2, 4.3, and 4.4 – point to model-free learning rates α_+^{MF} and α_-^{MF} that are lower than the model-based learning rates, and to a weight w on the model-based system that is fairly low – well below 0.5.

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, we find that the parameters of the framework that best fit the data put higher weight on the model-free system than on the model-based system. It is striking that investors would put a lot of 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, play a key role in explaining 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 4.7, 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 generates 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 different courses 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; and can help explain the dispersion across investors in their stock market allocations as well as the inertia in these allocations over time. 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 values of different courses 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

$$(1-\gamma)E(\log W_1) + \gamma(1-\gamma)E(\log W_2) + \gamma^2(1-\gamma)E(\log W_3) + \dots \tag{26}$$

where

$$W_t = W_0 \Pi_{\tau=1}^{t} R_{p,\tau} \tag{27}$$

is his wealth at time t. Substituting (27) into (26) and rearranging, the objective function becomes

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

as in (9).

Analytical solution to the portfolio problem (Section 2.2)

The value function $V(W_t)$ satisfies

$$V(W_t) = \max_{a_t} \left[(1-\gamma)E(\log W_{t+1}) + \gamma E_t(V(W_{t+1}))\right]. \tag{28}$$

We conjecture that the value function takes the form

$$V(W_t) = \log W_t + v, \tag{29}$$

where v is a constant that depends on the properties of the risky asset. Substituting (29) into (28), we obtain

$$v = \frac{1}{1-\gamma} \max_{a_t} E_t \log((1-a_t)R_f + a_t R_{m,t+1}).$$

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.

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.

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.

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.

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.

Hartzmark, S., Hirschman, S., and A. Imas (2020), "Ownership, Learning, and Beliefs," *Quarterly Journal of Economics*, forthcoming.

Jin, L. and P. Sui (2021), "Asset Pricing with Return Extrapolation," Working paper, Caltech.

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.

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.

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

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

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. We set $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, $\beta = 50$, $\gamma = 0.9$, $b = 0.0577$, $\mu = 0.01$, 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	**0.005**
10%	0	0	0	0	0.003	0.004
20%	0	0	0.004	0.004	**0.014**	0.014
30%	0	0	**0.016**	0.016	0.019	0.019
40%	0	0	0.004	0.004	0.004	0.004
50%	0	0	0	**0.002**	0.002	0.002
60%	0	0	0	0	0	0
70%	0	0	0	0	0	0
80%	0	-0.011	-0.011	-0.011	-0.011	-0.011
90%	0	**-0.051**	-0.051	-0.051	-0.051	-0.051
100%	0	-0.011	-0.011	-0.011	-0.011	-0.011

MODEL-BASED						
date	0	1	2	3	4	5
net market return		-17.4%	18.3%	-1.3%	12.8%	-16.6%
0%	0.011	0	0.076	0.03	0.334	0
10%	0.011	-0.005	0.078	0.031	0.338	-0.002
20%	0.012	**-0.01**	0.08	**0.032**	0.343	-0.004
30%	0.012	-0.015	0.081	0.033	0.347	-0.006
40%	0.012	-0.021	0.083	0.033	0.351	-0.009
50%	0.011	-0.027	0.084	0.033	0.355	-0.011
60%	0.011	-0.033	**0.084**	0.033	0.359	-0.014
70%	0.01	-0.04	0.085	0.033	**0.362**	-0.018
80%	0.009	-0.047	0.085	0.033	0.365	-0.021
90%	0.007	-0.054	0.084	0.032	0.368	-0.025
100%	0.005	-0.061	0.084	0.031	0.371	**-0.029**

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 mean model-free learning rates $\bar{\alpha}_+^{MF}$ and $\bar{\alpha}_-^{MF}$, mean model-based learning rates $\bar{\alpha}_+^{MB}$ and $\bar{\alpha}_-^{MB}$, and weight w on the model-based system. For comparison, the fifth column of the table reports the sensitivity when only the model-based system is used to make decisions. For mean learning rates of 0.1, 0.3, and 0.5, the rates are drawn from the uniform distributions $[0.05, 0.15]$, $[0.1, 0.5]$, and $[0.3, 0.7]$, respectively. There are 300,000 investors: six cohorts of 50,000 investors each which enter financial markets at different dates. We set $L = T = 30$, $\beta = 50$, $\gamma = 0.9$, $b = 0.0577$, $\mu = 0.01$, and $\sigma = 0.2$.

$\bar{\alpha}_+^{MF}$, $\bar{\alpha}_-^{MF}$	$\bar{\alpha}_+^{MB}$, $\bar{\alpha}_-^{MB}$	w	Action vs. belief	Model-based benchmark
0.3	0.3	0.2	0.98	2.58
0.1	0.3	0.2	1.0	2.58
0.5	0.3	0.2	0.96	2.58
0.3	0.1	0.2	1.5	3.47
0.3	0.5	0.2	0.79	2.3
0.3	0.3	0	0.45	2.58
0.3	0.3	0.5	1.82	2.58

Table 3. Individual and time fixed effects in investor beliefs. The table reports the percentage of the variation in investor beliefs that is explained by an individual fixed effect (FE) alone; by a time fixed effect alone; and by both fixed effects together. The first four rows of numbers correspond to different cross-sectional distributions of model-based learning rates α_+^{MB} and α_-^{MB}. There are 3,000 investors: six cohort of 500 investors each. We set $L = T = 30$, $\beta = 50$, $\gamma = 0.9$, $b = 0.0577$, $w = 0.2$, $\mu = 0.01$, and $\sigma = 0.2$. The fifth row reports the empirical estimates of Giglio et al. (2021).

α_+^{MB}, α_-^{MB}	Individual FE	Time FE	Both FE
(0.05,0.15)	72	27	99
(0.05,0.35)	53	40	93
(0.1,0.5)	40	50	90
(0.05,0.55)	50	50	100
Empirical	*48*	*5*	*52*

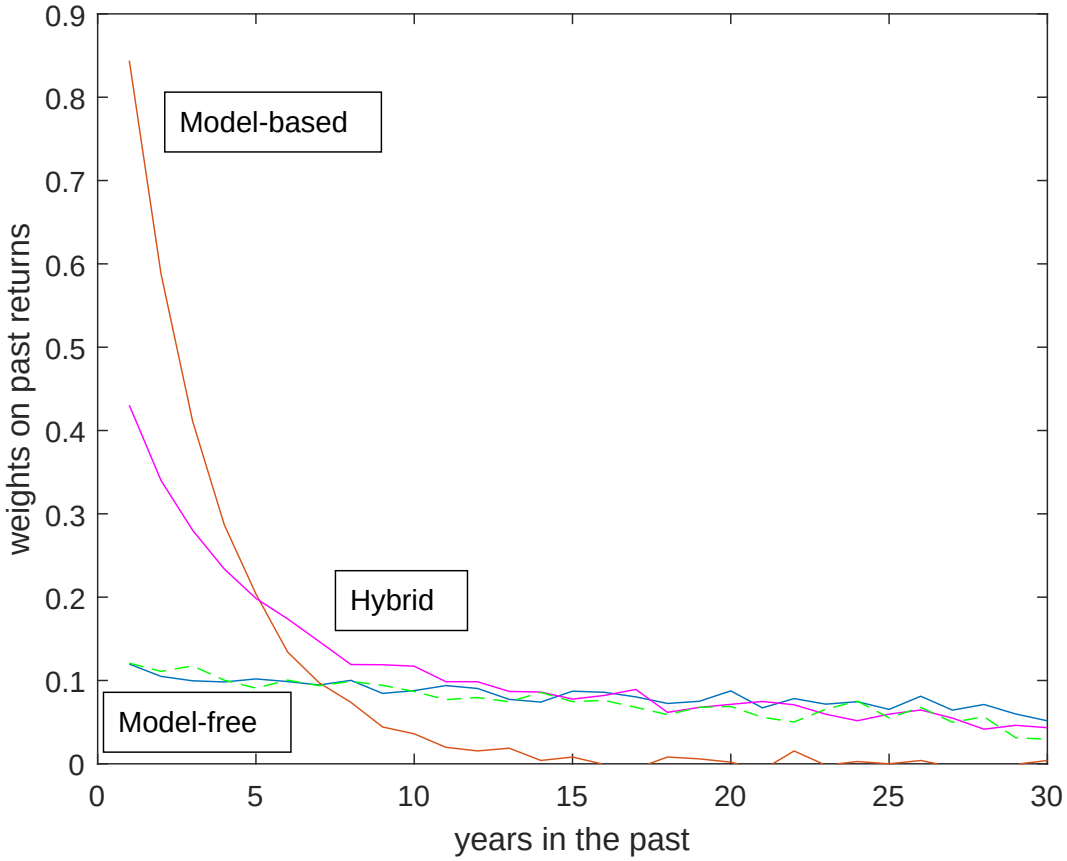


Figure 1. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ and plot the coefficients for four cases: a model-free system, a model-based system, and a hybrid system (solid lines); and a "pure" model-free system without generalization (dashed line). There are 50,000 investors. We set $L = T = 30$, $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, $\beta = 50$, $\gamma = 0.9$, $w = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$ except in the case of the pure model-free system.

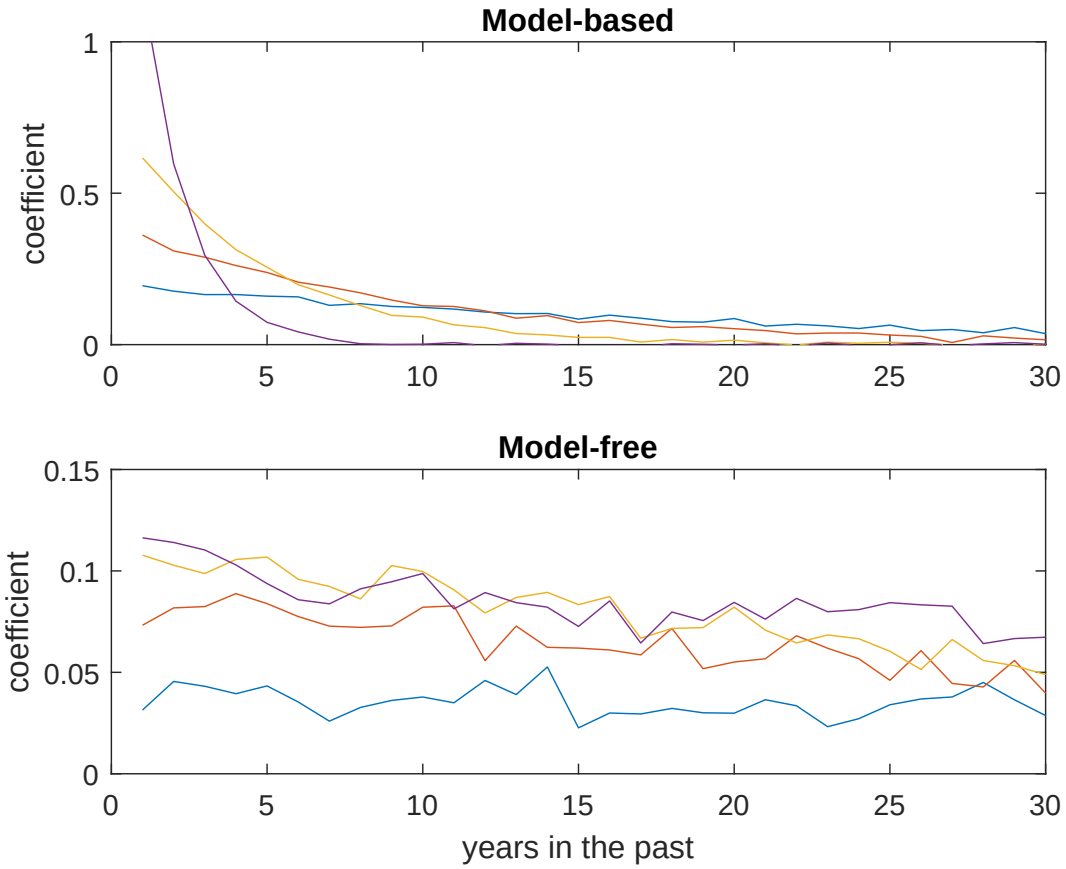


Figure 2. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ for two cases: a model-free system and a model-based system. The top graph plots the coefficients for the model-based system for four different values of the model-based learning rates α_+^{MB} and α_-^{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 model-free learning rates α_+^{MF} and α_-^{MF}, namely 0.05 (blue), 0.1 (red), 0.2 (yellow), and 0.5 (magenta). There are 50,000 investors. We set $L = T = 30$, $\beta = 50$, $\gamma = 0.9$, $w = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

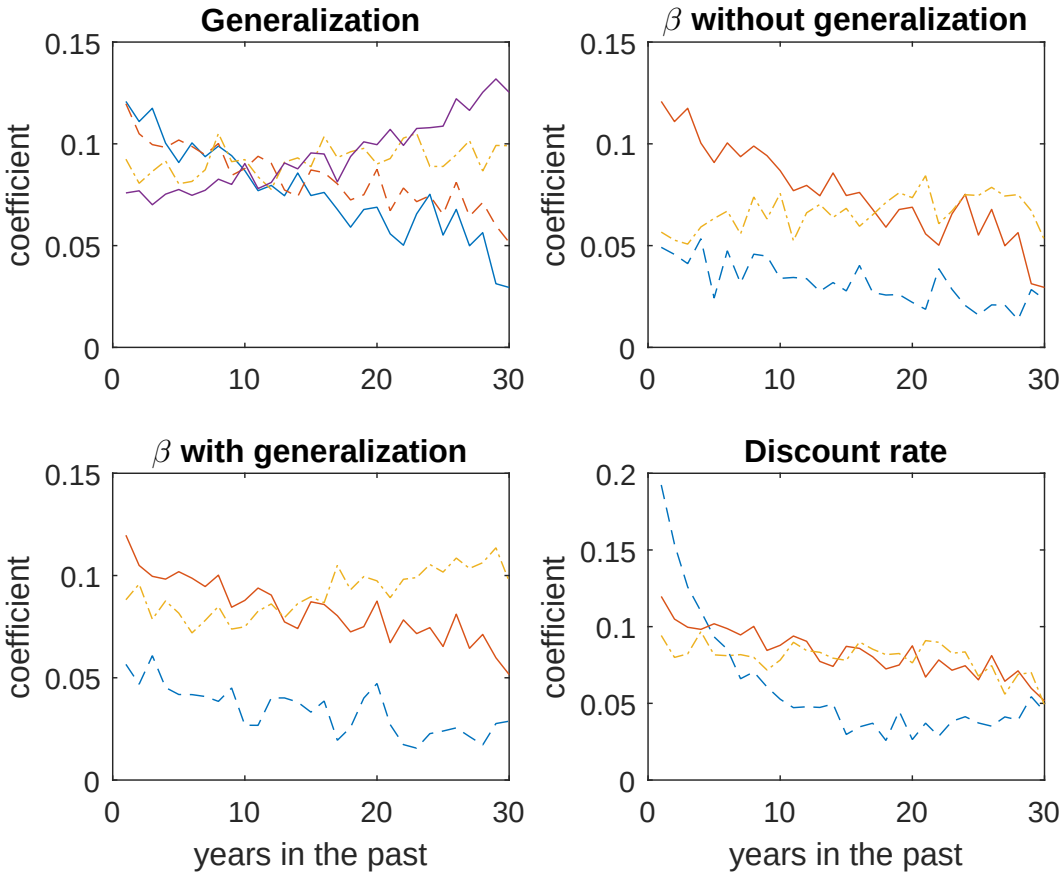


Figure 3. We run regressions of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T-j}\}_{j=0}^{j=29}$ for the model-free system and for different sets of parameter values. The lines in the top-left graph correspond to four values of the generalization parameter b, namely 0.001 (blue), 0.0577 (red), 0.115 (yellow), and 0.23 (magenta). The lines in the top-right and bottom-left graphs correspond to three values of the exploration parameter β, namely 10 (blue), 50 (red), and 500 (yellow); the top-right graph removes generalization, so that $b = 0.001$, while the bottom-left graph allows for it with $b = 0.0577$. The lines in the bottom-right graph correspond to three values of the discount rate γ, namely 0.3 (blue), 0.9 (red), and 0.99 (yellow). There are 50,000 investors. For the other parameters, we set $L = T = 30$, $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, $\beta = 50$, $\gamma = 0.9$, $w = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

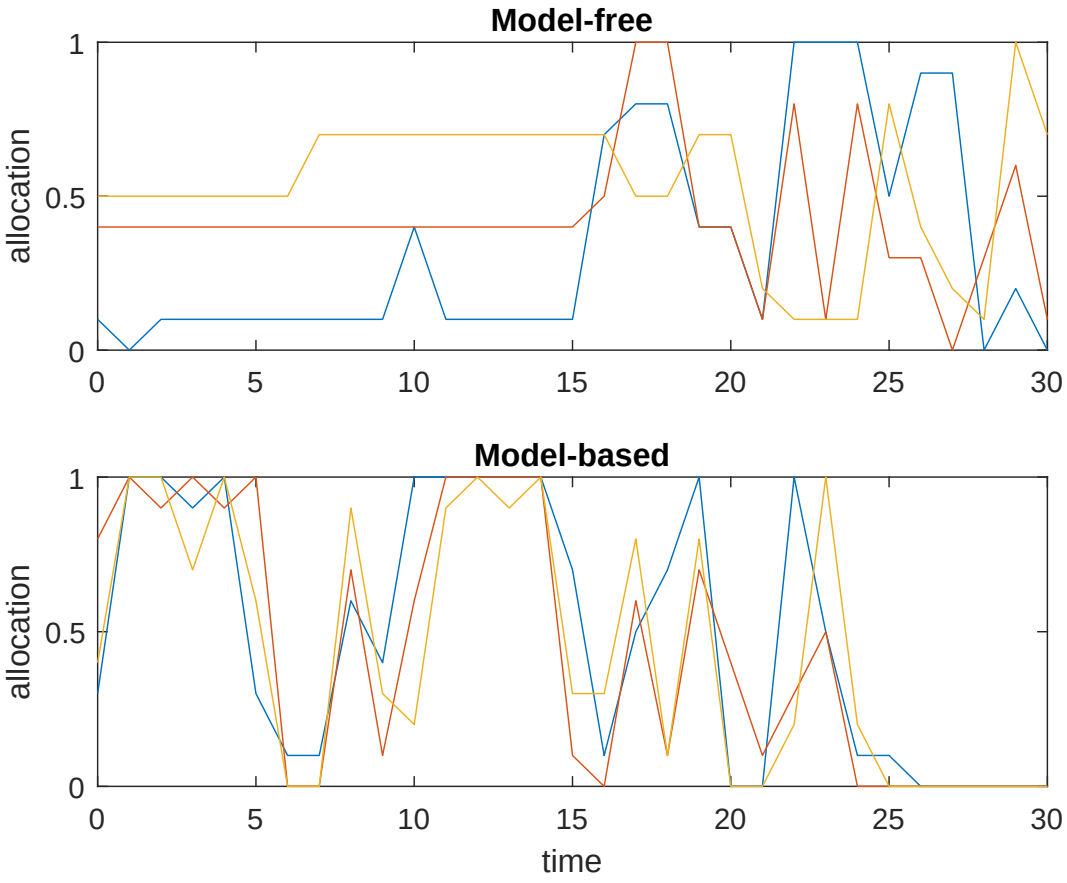


Figure 4. The top panel plots the stock market allocations over time of three investors who use only the model-free system to make decisions. The bottom panel plots the allocations of three investors who use only the model-based system to make decisions. The parameter values are $L = T = 30$, $\alpha_+^{MF} = \alpha_-^{MF} = \alpha_+^{MB} = \alpha_-^{MB} = 0.3$, $\beta = 200$, $\gamma = 0.9$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

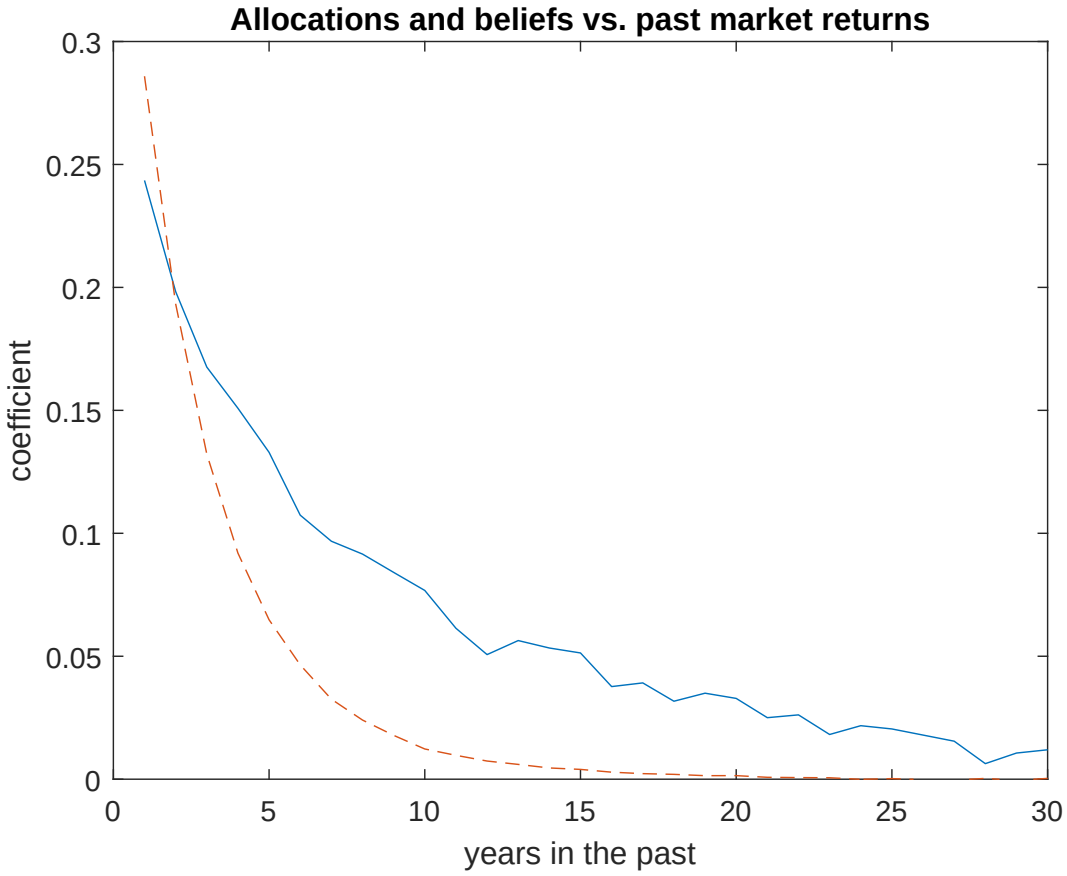


Figure 5. The solid line plots the coefficients in a regression of the stock market allocation a_T at date T chosen by investors who use a hybrid system to make decisions on the past 30 years of stock market returns the investors were exposed to. The dashed line plots the coefficients in a regression of investors' expectations at time T about the future one-year stock market return on the past stock market returns. There are 300,000 investors: six cohorts of 50,000 investors each who enter financial markets at different times. For each investor, each of α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} is drawn independently from the uniform distribution $[0.1, 0.5]$. We also set $L = T = 30$, $\beta = 50$, $\gamma = 0.9$, $b = 0.0577$, $w = 0.2$, $\mu = 0.01$, and $\sigma = 0.2$.

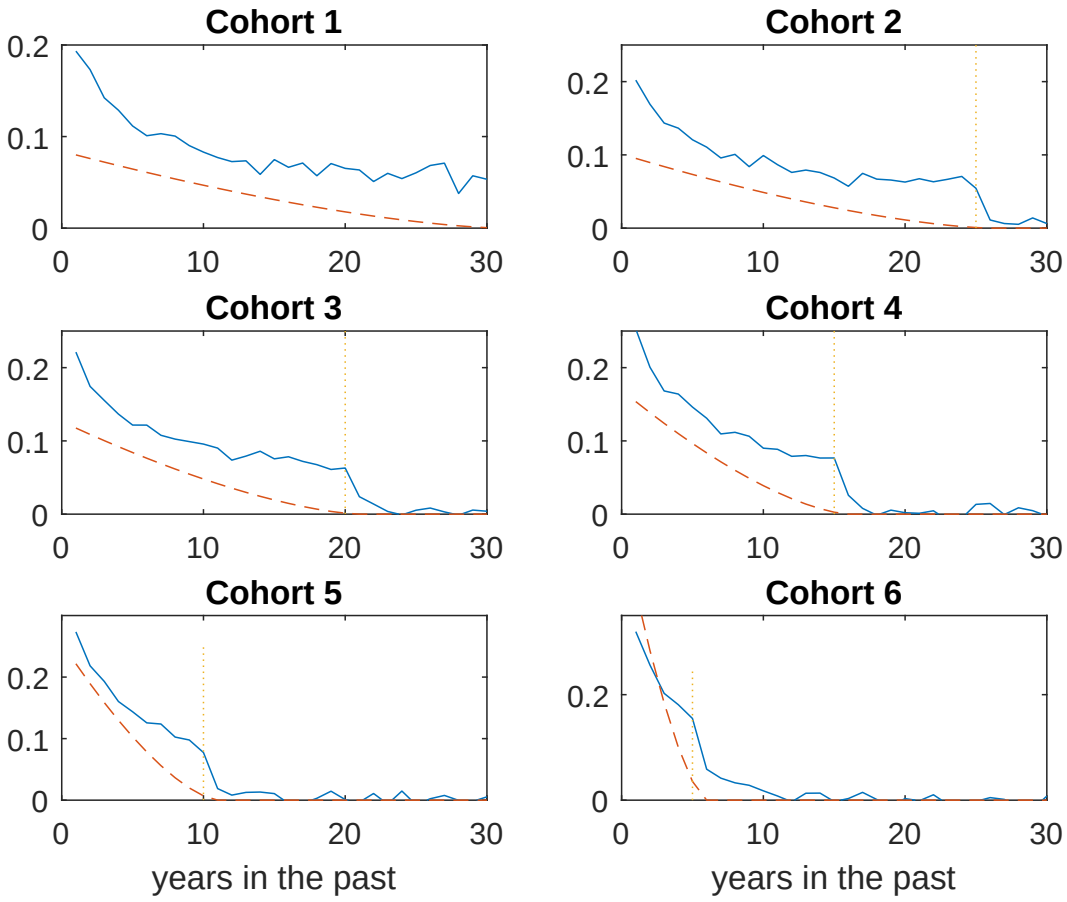


Figure 6. The six graphs correspond to six cohorts of investors. In each graph, the solid line plots the coefficients in a regression of the time T stock market allocations a_T of the investors in that cohort on the past returns they have been exposed to. The six cohorts have different numbers of years 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 α_+^{MF}, α_-^{MF}, α_+^{MB}, and α_-^{MB} is drawn independently from the uniform distribution $[0.1, 0.5]$. We also set $L = T = 30$, $\beta = 50$, $\gamma = 0.9$, $b = 0.0577$, $w = 0.2$, $\mu = 0.01$, 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 $(n - k)^\lambda / \Delta$, where k is the number of years in the past, $\lambda = 0.15$, and Δ is a normalizing constant.