INTERNET APPENDIX

A. Experimental Evidence of Model-free Behavior

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

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

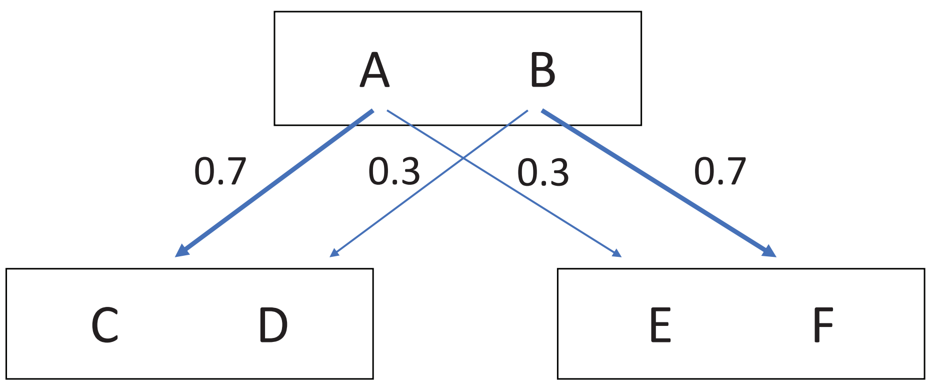

Figure A1. The diagram shows the structure of an experiment in Daw et al. (2011). In the first stage, the participant has a choice between two options, A and B; in the second stage, he chooses either between options C and D or between options E and F. The arrows indicate the transition probabilities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

The model-free and model-based systems make different predictions about behavior in this setting. Suppose that the individual chooses A in the first stage and is then offered a choice between E and F; suppose that he chooses E and then receives a reward. Under the model-free system, he will be inclined to choose A again in the next trial because this choice was ultimately rewarded. Under the model-based system, however, he will be inclined to choose B in the next trial: the model-based system makes use of information about the

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

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

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

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

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

B. Time-varying Weights on the Two Systems

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

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

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

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

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

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

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

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

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

The transition rates are functions of system reliability:

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

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

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

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

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

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

where

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

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

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

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

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

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

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

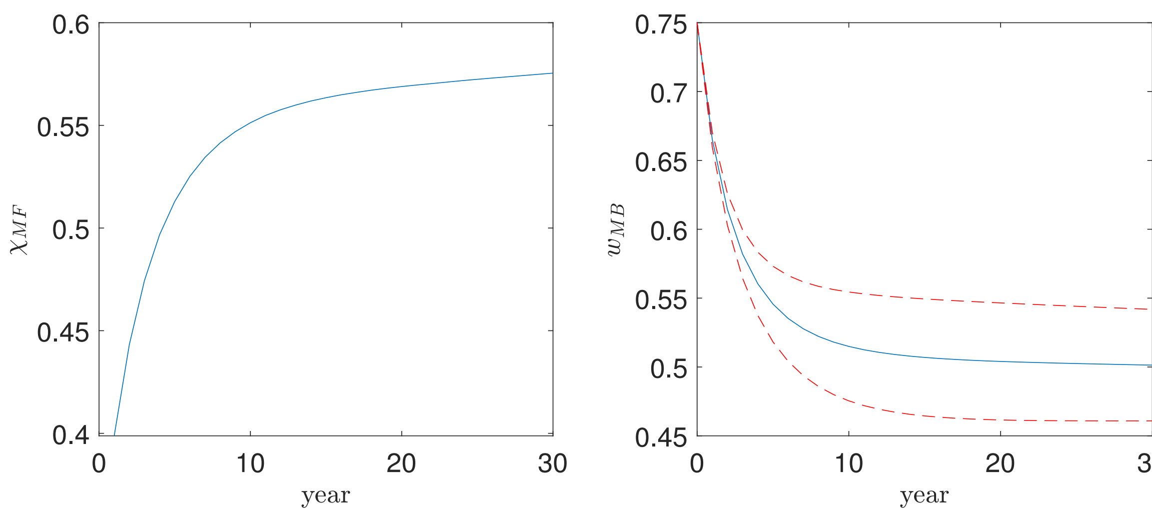

Figure A2. The left graph shows the evolution, over time, of the average investor measure of the reliability of the model-free system. The solid line in the right graph shows the mean weight that investors put on the model-based system. The dashed lines plot the weights on the model-based system that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

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

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

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

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

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

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

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

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

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

Generalization. The top-left graph in Figure 3 plots the coefficients in a regression of the time T model-free allocation on past stock market returns for four values of the generalization parameter b: 0, 0.0577, 0.115, and 0.23. The first of these values corresponds to no generalization; the other three values give the Gaussian function in equation (14) of the main text, normalized as a probability distribution, a standard deviation equal to that of a uniform distribution whose support has a width of 0.2, 0.4, and 0.8, respectively.

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

Exploration. The top-right graph in Figure 3 plots the relationship between the model-free allocation and past market returns for three different values of β, which controls the degree of exploration, namely 10, 50, and ∞. Recall that, as β rises, the investor explores less: he is more likely to choose the allocation with the highest estimated Q value; when $\beta = \infty$, he always chooses this allocation. We find that, for a wide range of values of β – any β below 80 – the model-free allocation puts positive weights on past returns that decline over most of the time range, as they do for our benchmark case of $\beta = 30$. However, when β is higher than 80, the weights on past returns increase for more than half of the time range. To see why, suppose that, soon after the investor enters financial markets, the stock market posts a high return, raising the Q value of his most recent allocation. If the value of β is high, the investor is likely to stick with this allocation for a substantial period of time. As such, the early returns he experiences have a large effect on his subsequent allocations.

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

Number of allocations. In the main text, we allowed investors to select from one of 11 possible allocations. The bottom-right graph in Figure 3 shows how the time T model-free allocation depends on past market returns as we vary the number of allocation options, ranging from three, namely {0%, 50%, 100%}, up to 21, namely {0%, 5%, ... , 95%, 100%}. The graph shows that, as we lower the number of possible allocations, the relationship between the time T allocation and past returns, while initially downward-sloping, becomes much flatter, thereby giving distant past returns a larger role. This property of the model-free system again distinguishes it from the model-based system, where the number of possible allocations has little impact on the relationship between the time T allocation and past returns.

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

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

D. Analytical Results

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

We start with the case in which the learning rates $\alpha = 1$, the discount factor $\gamma = 0$, and there are just two possible allocations, namely $a = 0$ and $a = 1$. For model-free and model-based learning, respectively, Theorems 1 and 2 below present analytical results on how the allocation recommended by each system depends on past market returns. By comparing equations (9) and (18), we confirm that the model-free system puts substantially greater weight on distant past returns.

We then turn to the less restrictive case where the learning rates α can take any value between 0 and 1; once again, $\gamma = 0$ and there are two possible allocations. For the model-free and model-based algorithms, respectively, Theorems 3 and 4 below present analytical results about the dependence of the recommended allocation on past market returns. Comparing equations (21) and (34)-(35), we again see that the model-free algorithm puts substantially greater weight on distant past returns.

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

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

Given these assumptions, the following result holds:

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

for $k \geq 0$.

Proof: At any time $t > 0$,

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

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

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

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

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

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

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

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

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

Equation (12) allows us to derive $\frac{\partial \mathbb{E}[a_t]}{\partial R_{m,t-k}}$, the sensitivity of the expected allocation to past returns. We first consider the case with $k = 0$. In this case,

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

As t goes to infinity, we obtain

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

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

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

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

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

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

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

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

Given these assumptions, the following result holds:

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

Proof: At any time $t > 0$,

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

The softmax rule implies

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

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

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

Given these assumptions, the following result holds:

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

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

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

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

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

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

To further derive $\frac{\partial e^{\beta Q_t(1)}}{\partial R_{t-k}}$ in (25), we let n denote the number of indices i, with $i \in \{t-k, \ldots, t-1\}$ and $b_i = 1$. We then proceed by considering two cases. The first case is when $b_0 = b_1 = \ldots = b_{t-k-2} = 0$. In this case, $Q_t(1)$ can be written as the sum of $(1-\alpha)^n \log(R_{t-k})$ and a term unrelated to R_{t-k}. As such,

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

and (25) can be simplified as

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

The second case is when $b_0, \ldots, b_{t-k-2}$ are not all equal to zero. In this case, $Q_t(1)$ can be written as the sum of $\alpha(1-\alpha)^n \log(R_{t-k})$ and a term unrelated to R_{t-k}. As such, (25) can be simplified as

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

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

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

Note that

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

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

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

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

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

We now turn to (24). We have

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

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

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

Given these assumptions,

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

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

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

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

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

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

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

and for $k = t-1$,

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

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

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

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

E. Parameter Estimation

In this section, we describe in more detail the procedure that we use in Section 4.7 to estimate the values of four important parameters in our framework: the mean model-based learning rate across investors $\bar{\alpha}^{MB}$; the mean model-free learning rate $\bar{\alpha}^{MF}$; the exploration parameter β; and the weight w on the model-based system. We do the estimation in two steps. We first use data on investor beliefs to estimate $\bar{\alpha}^{MB}$. We then estimate $\bar{\alpha}^{MF}$, β, and w by targeting two facts discussed in Section 4 of the paper, namely the sensitivity of allocations to beliefs in Giglio et al. (2021) and the experience effect in Malmendier and Nagel (2011). We keep the remaining parameters at their benchmark values, namely $L = T = 30$, $\gamma = 0.97$, $\Delta = 0.5$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0.0577$.

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

To do this, we take $300,000$ investors in six cohorts of $50,000$ each; each investor sees a different sequence of stock market returns from time $t = -L$ to time $t = T$. For a given value of $\bar{\alpha}^{MB}$, we draw each investor's model-based learning rates, α_+^{MB} and α_-^{MB}, from a uniform distribution centered at $\bar{\alpha}_{MB}$ and with width $\Delta = 0.5$. We then compute investor beliefs at each time, as determined by the model-based system and in particular by equations (17) and (18) in the main text. Finally, we regress investors' beliefs at time T on the past 30 years of stock market returns they have been exposed to, and record the coefficients c_1, c_2, and c_3 on the annual returns one, two, and three years in the past, respectively. We repeat this exercise for many different values of $\bar{\alpha}^{MB}$ and select the value of $\bar{\alpha}^{MB}$ that minimizes

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

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

With this value of $\bar{\alpha}^{MB}$ in hand, we search for values of $\bar{\alpha}^{MF}$, β, and w that best match two empirical targets. The first is the coefficient in a regression of investor allocations on investor beliefs, which Giglio et al. (2021) find to be approximately 1. For given values of $\bar{\alpha}^{MF}$, β, and w, we can compute this coefficient, d, in simulated data from our framework. Due to computational constraints, these data are now based on 60,000 investors in six cohorts of 10,000 each.

Our second target is the functional form in (28) in the main text with $\lambda = 1.3$, which Malmendier and Nagel (2011) use to capture empirical experience effects. Intuitively, we are looking for parameter values that minimize the distance between unnormalized versions of the solid and the dashed lines in the six graphs in Figure 6. For given values of $\bar{\alpha}^{MF}$, β, and w, and for cohort 1, we run a regression in our simulated data of the time T allocations on the past 30 years of returns. We then compute another vector of 30 coefficients given by

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

which, according to column (i) in Table IV of Malmendier and Nagel (2011), captures the empirical relationship between allocations and returns j years in the past for a cohort of age 30. We then compute the L^2 norm of the difference between the two vectors and call this

MSE_1, the mean-squared error for cohort 1. In a similar way, we compute MSE_i for $i = 2$ to 6, which correspond to cohorts 2 through 6.

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

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

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

F. Allowing for State Dependence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The hybrid Q values are

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

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

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

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

We now present some numerical analysis. The parameters σ, $\alpha_{t,\pm}^{MF}$, $\alpha_{t,\pm}^{MB}$, γ, w, and β take the baseline values used in Figure 1 of the paper. In addition, we set $\theta = 0.8$, $\mu_H = 6\%$, $\mu_M = 1\%$, $\mu_L = -4\%$, $\overline{S} = \exp(\mu_M + 0.5\sigma^2) + 3\% = 1.0605$, and $\underline{S} = \exp(\mu_M + 0.5\sigma^2) - 3\% = 1.0005$. The simulation setup is the same as in Figure 1; in particular, there are 300,000 investors. We consider two cases: the case where investors do not recognize the existence of the three states, and the case where they do recognize and observe the three states. In each case, we study the performance and recommended allocations of the model-free system, the model-based system, and the hybrid system. To evaluate performance, we look at each investor's excess portfolio return from t to $t+1$, where t goes from 0 to 29; we compute the mean and standard deviation of these 30 excess returns for each investor; finally, we average these numbers across the 300,000 investors. To study allocations, we look at each investor's portfolio allocation at time 30; we then average these allocations across the investors who are facing an asset that is in state s at time 30, where s is L, M, or H.

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

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

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

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

We make three observations about these results.

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

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

When the model-based system is able to recognize the three market states but the model-free system is not, a tension arises between the two systems, as suggested in Section 5.2. Following a sequence of good returns, the model-free system recommends a high allocation: when $S_{m,t} > \overline{S}$, the average allocation recommended by the state-independent model-free system is 60.53%, higher than what it recommends in the Moderate state. By contrast, the model-based system recognizes that a good trend is often followed by low returns and hence recommends a low allocation: when $S_{m,t} > \overline{S}$, the average allocation recommended by the state-dependent model-based system is 41.97%, lower than what it recommends in the Moderate state. One system therefore pulls the investor toward a higher allocation in the Low state, while the other pulls him toward a lower allocation.

G. SARSA: An Alternative Model-free Framework

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

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

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

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

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

Analogous to the analysis in Section 3.2, we examine how investors' date T allocations a_T recommended by each of SARSA, Q-learning, and model-based learning depend on the past market returns investors have been exposed to. Figure A3 presents the results and leads to two observations. First, for SARSA and Q-learning, the weights the allocation a_T puts on past stock market returns are quantitatively similar. The only exception is the weight on the most recent stock market return: in the case of SARSA, the allocation a_T is determined by Q values that do not depend on the most recent return $R_{m,T}$; this allocation therefore puts zero weight on $R_{m,T}$. Second, while the allocation recommended by model-based learning depends primarily on recent past returns, the allocations recommended by both Q-learning and SARSA depend significantly even on distant past returns.

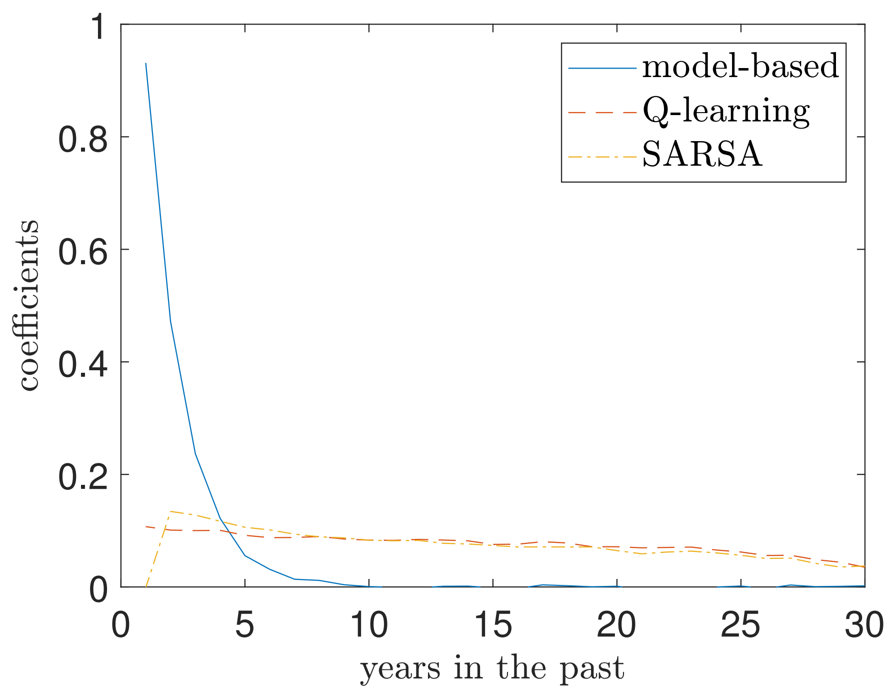

Figure A3. We run a regression of investors' allocations to the stock market a_T at time T on the past 30 years of stock market returns $\{R_{m,T+1-j}\}_{j=1}^{30}$ investors were exposed to and plot the coefficients for three cases: model-based learning; model-free Q-learning; and model-free SARSA. There are 300,000 investors. We set $L = T = 30$, $\alpha_{\pm}^{MF} = \alpha_{\pm}^{MB} = 0.5$, $\beta = 30$, $\gamma = 0.97$, $\mu = 0.01$, $\sigma = 0.2$, and $b = 0$, so that there is no generalization.