PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Data-driven equation discovery reveals nonlinear reinforcement learning in humans

Kyle J. LaFollette[a,b,1], Janni Yuval[c,2], Roey Schurr[d], David Melnikoff[e], and Amit Goldenberg[d,f,g]

Affiliations are included on p. 10.

Edited by James McClelland, Stanford University, Stanford, CA; received July 11, 2024; accepted June 24, 2025

Computational models of reinforcement learning (RL) have significantly contributed to our understanding of human behavior and decision-making. Traditional RL models, however, often adopt a linear approach to updating reward expectations, potentially oversimplifying the nuanced relationship between human behavior and rewards. To address these challenges and explore models of RL, we utilized a method of model discovery using equation discovery algorithms. This method, currently used mainly in physics and biology, attempts to capture data by proposing a differential equation from an array of suggested linear and nonlinear functions. Using this method, we were able to identify a model of RL which we termed the Quadratic Q-Weighted model. The model suggests that reward prediction errors obey nonlinear dynamics and exhibit negativity biases, resulting in an underweighting of reward when expectations are low, and an overweighting of the absence of reward when expectations are high. We tested the generalizability of our model by comparing it to classical models used in nine published studies. Our model surpassed traditional models in predictive accuracy across eight out of these nine published datasets, demonstrating not only its generalizability but also its potential to offer insights into the complexities of human learning. This work showcases the integration of a behavioral task with advanced computational methodologies as a potent strategy for uncovering the intricate patterns of human cognition, marking a significant step forward in the development of computational models that are both interpretable and broadly applicable.

reinforcement learning | dynamical systems | nonlinear modeling | machine learning

Significance

Our article offers an answer to a foundational question in psychology and neuroscience: how do people learn from rewards and punishments? Specifically, we introduce a computational model of human reinforcement learning (RL) that points to a nonlinear updating of the probability of reward. The strength of our model lies also in the process through which it was developed. Specifically, we discovered our model in a bottom–up fashion using symbolic regression—a class of machine learning tools applied primarily in physics and engineering. We believe that, in addition to the theoretical contributions of the model to the field of RL, our work strongly demonstrates the utility of implementing equation-discovery tools in the field of social behavior.

Over the past few decades, the social sciences have seen an increasing prevalence of computational cognitive modeling for explaining human behavior (1). Computational models have had a transformational contribution to a variety of domains, most notably reinforcement learning (RL) (2). RL provides a mathematical framework for understanding how agents learn and make decisions based on experience with rewards or punishments. Research on RL has contributed to our understanding of human and animal learning, including its neuronal underpinnings in the brain (3–7). Insights from RL in the social sciences have also been adopted in machine learning, contributing to tremendous improvements in facilitating learning in artificial agents (8–10).

Although undoubtedly successful, RL models traditionally update reward expectations linearly, an assumption that may oversimplify human behavior's complex relationship with rewards. Contrary to this linear approach, evidence outside of RL models suggests that human behavior exhibits a nonlinear response to rewards, with subjective value not scaling linearly with the reward's objective size. This is supported by both psychological and economic theories (11–13), as well as neuroscientific findings, indicating a nonlinear coding of rewards in the brain (14–17). One of the most studied aspects of this nonlinearity is probability weighting, a concept central to decision-making models such as Cumulative Prospect Theory (CPT; 13). CPT proposes an inverse "S-shaped" weighting function, in which low probabilities are overweighted and high probabilities are underweighted. However, this is only one of many proposed functional forms in behavioral economics. Alternatives,including those by Prelec (14), Gonzalez and Wu (15), and others,have emphasized different curvature properties, parametric flexibility, and psychological interpretations. Moreover, empirical findings suggest that these distortions differ systematically between decision-from-description tasks (used in most CPT work) and decision-from-experience tasks. The latter often shows reversed or flattened weighting patterns (e.g., underweighting of rare events), raising questions about the stability of these effects across contexts (16–18).

Despite the empirical evidence for nonlinearities in valuation and utility, most RL models continue to use linear delta-updating rules for learning. The Rescorla–Wagner

Author contributions: K.J.L., J.Y., R.S., D.M., and A.G. designed research; K.J.L. and A.G. performed research; K.J.L. analyzed data; and K.J.L., J.Y., R.S., D.M., and A.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: kyle.lafollette@chicagobooth.edu.

[2]Present address: Google Research, Mountain View, CA 94043.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2413441122/-/DCSupplemental.

Published July 31, 2025.

model (19), arguably one of the most influential delta-updating rules, assumes exactly this: A linear relationship between expectations and change in response to feedback. Variations of the model that incorporate decay over time, or asymmetric learning rates for positive and negative feedback, all share this common assumption that learning is linear. Prior work has incorporated nonlinear transformations of outcome values,such as risk-sensitive utilities or probability weighting functions (e.g., refs. 20–22), however these typically modify the inputs to the prediction error, leaving the structure of the learning rule itself unchanged. This distinction between nonlinear inputs and nonlinear updating has important implications for how models capture learning dynamics and the emergence of systematic biases.

This underscores a complex problem in model development: despite their achievements, RL models—and computational models of social behavior in general—are vulnerable to the biases and limitations of their designers, as they are mostly developed top–down based on theoretical insights or adapted from historically dominant models. This is perhaps why canonical RL models struggle to find a balance between interpretability, parsimony, accuracy, and generalizability across individuals and contexts (23). New models are being proposed continuously; however, they suffer from many of the same limitations as the models they aim to replace (24).

Deep learning may come to mind as a suitable alternative to top–down model development, but comes with its own tradeoffs: high prediction accuracy at the expense of interpretability and limited generalization outside training data. Recently, however, efforts to merge deep learning with traditional models have aimed at enhancing interpretability and systematic discovery (25–28). Constraining deep learning within the bounds of theory has yielded more understandable models (29–32), though their broad applicability remains unproven. A complementary approach is to improve existing interpretable models using bottom–up, machine learning, approaches (33). These methods, while innovative, still depend on preexisting models and extensive data. To address these gaps and promote model discovery in the social sciences, we propose to adopt algorithms designed for data-driven discovery of nonlinear differential equations in physics and engineering. These data-driven approaches allow the freedom to explore a vast range of functional forms in relatively small datasets while constraining the models to be interpretable.

The notion that dynamic models can be discovered using bottom up approaches received increased attention in recent decades, especially in physics (34, 35). Early work in this space suffered from overfitting and required immense computing power. However, recent developments allow for implementations of bottom–up equation discovery in complex, noisy, and multidimensional systems (36–38), making it well suited for model discovery in social sciences. Unlike other, more opaque machine learning approaches, these algorithms generate systems of equations that researchers can interpret. Users can also predetermine the space of possible terms that describe the system and control the level of complexity of the obtained model.

Here, we utilize an equation discovery algorithm, SINDy (Sparse Identification of Nonlinear Dynamics; 24), to develop and improve human RL models. SINDy is based on the idea of sparse regression, seeking to identify a minimal set of ordinary differential equations that aim to describe the underlying dynamics of a system that produced the observed data (here, the underlying cognitive process). It uses a combination of optimization and feature selection to find the sparse set of candidate functions through iterative multiple regression, and it can amalgamate a wide variety of linear and nonlinear terms (see *Methods* for details). SINDy has been applied in physics (39, 40), engineering (41, 42), and biology (24, 43). An introductory paper suggested its use in social sciences (44), but it has not been used yet for model development with empirical data.

The goal of the current project is to discover models of RL. We use SINDy to enable testing of multiple RL models without the biases inherent to traditional top–down model development. In phase 1, we designed a simple RL task that allows us to capture participants' estimation of a probability of reward across multiple trials. Using SINDy, we then revealed a model—termed the Quadratic Q-Weighted model—that introduces unique behavioral insights into how people learn the probability of reward. This model, in line with probability weighting theories, demonstrates that participants exhibit a systematic distortion in their estimation or probability, which is similar to the nonlinear probability weighting seen in previous decision-making research. What sets the model apart, however, is its ability to capture a dynamic transition between S-shaped and inverse S-shaped distortions, revealing a context-dependent flexibility influenced by participants' expectations. In phase 2 we then take the Quadratic Q-Weighted model and compare its ability to predict reward data on completely different kinds of tasks involving evaluating reward in much more complicated situations such as a two-armed bandit task. We do not use SINDy directly in this phase; rather, we take the Quadratic Q-Weighted model discovered using our simple RL task and embed that model within existing models of more complex decision-making. We demonstrate that the application of the Quadratic Q-Weighted model achieves better results than previous state-of-the-art models across eight of nine public datasets, each published in leading academic journals. This work therefore makes a two-fold contribution: first, it provides a proof of concept for utilizing an equation discovery algorithm in the social sciences, enabling the discovery of a RL model from behavioral data. Second, it introduces a model of human RL that accounts for probability weighting distortions and demonstrates its generalization capabilities to more complex decision-making tasks, thereby unveiling insights into human cognition.

Results

Phase 1: Equation Discovery from Empirical Probability Estimates. Our first goal was to determine whether algorithms discovered by SINDy can provide insights into probabilistic learning when trained on empirical data from human learners. To this end, we conducted two empirical studies using a learning task composed of 100 trials. Participants assumed the role of an inspector tasked with identifying the rate at which a factory produces working versus defective phones (Fig. 1). On each of the 100 trials, participants inspected a new phone produced by the factory and learned whether it was working or defected. Participants then reported the probability that the next phone would be working (see *Methods* for detailed description). The true probability of receiving a working phone changed trial-to-trial according to a Gaussian random walk (SD = 0.1), bounded between 0.1 and 0.9; the initial value was drawn from a uniform distribution in the range 0.1 to 0.9. To incentivize accurate predictions, we offered participants a $0.03 bonus per response within 5% of the true probability. Attention checks were included during the task to ensure data quality; participants who did not pass our criteria for attention checks were excluded from analysis (*Methods*).

We ran two versions of the task. In Study 1 (N = 455), we set the initial probability of a working phone to 0.5. This probability changed every trial according to a Gaussian random walk with

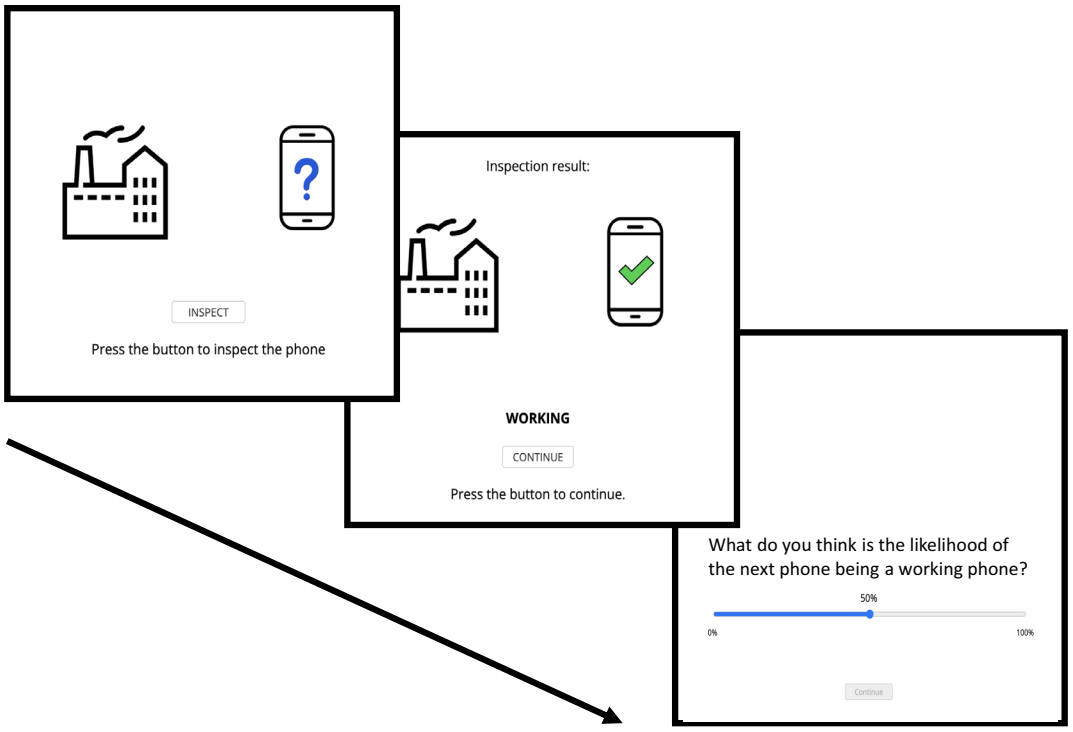

Fig. 1. Structure of learning task used in Studies 1 and 2. Participants inspected phones produced from an assembly line. On each trial, a single phone was revealed to either be working or defective. Following each observation, participants were asked to rate on a scale from 0 to 100% what they thought was the likelihood of the next phone being a working phone.

SD = 0.025; the random walk was unique for every participant. We used diffusion in the true reward rate in order to keep participants engaged with the task, as done in similar tasks (45–47).

In Study 2 (N = 177), the task was the same as in Study 1 except for two modifications. First, the initial value of the true probability of a working phone was randomly drawn from a uniform distribution $U(0.1, 0.9)$ rather than being fixed at 0.5. Second, we increased the SD of the random walk from 0.025 to 0.1. The purpose of these modifications was to explore how SINDy performed across a broad range of task parameters. Neither Studies 1 or 2 were preregistered and all analyses should be considered exploratory.

We trained SINDy using data from all participants who met our inclusion criteria (see *Methods* for exclusions), separately for each study. Input data provided to SINDy were limited to participants' reported expectations of observing a working phone Q_t, their observations of whether a phone was working or defective r_t, and trial number t. We also provided SINDy with a matrix of candidate functions for feature selection, allowing for a variety of models to be identified. These included identity functions for previous expectations and reward, time-dependent decaying functions, and exponential functions for nonlinearity (see *Methods* for specifics on candidate functions and fitting procedure). Consequentially, the Rescorla–Wagner model could be discovered by SINDy if it best explained the empirical data from either study. This was ensured through a series of simulation studies (*Simulations*).

For Study 1, SINDy discovered the following model (R^2 = 0.204):

$$Q_{t+1} = 0.11r_t - 0.24{Q_t}^2.$$

For Study 2, SINDy discovered a near identical model (R^2 = 0.196):

$$Q_{t+1} = 0.10r_t - 0.17{Q_t}^2.$$

To demonstrate that these models were superior in fit to the Rescorla–Wagner model, we separately trained SINDy with a smaller matrix of candidate features limited to only the r-Q term. This limited SINDy to only discover the Rescorla–Wagner model. These limitations yielded worse fit in both studies (Study 1 R^2 = 0.144; Study 2 R^2 = 0.174).

Note that the coefficients of the discovered models' parameters are not symbolic and are fixed across participants. In both studies, SINDy discovered models of identical form, albeit with slightly different numerical values. We termed the model that SINDy produced the Quadratic Q-Weighted model since the model includes a quadratic term on previous expectation rather than a linear one (hence "Quadratic") and the model includes unequal scaling coefficients for present reward and previous expectation (Q-value; hence "Q-Weighted"). The Quadratic Q-Weighted model accounts for several interesting behavioral phenomena discussed in the results. Most importantly, the functional form of the Quadratic Q-Weighted model leads to an asymptotic bias in the estimation of the true probability. Namely, the model implies that over the long term, participants tend to underestimate Q values when reward probability is high and tend to overestimate Q values and reward probability is low. The transition between under- and overestimation happens approximately when the true probability of reward is equal to a/b where

$$Q_{t+1} = ar_t - b{Q_t}^2.$$

Fig. 2 illustrates why the Quadratic Q-Weighted model implies such over/under estimation, showing the change in Q as a function of either reward or no-reward and as dependent on previous Q (see *SI Appendix* for a proof of this point of under-to-over estimation). For low values of Q, the change in Q in the Quadratic Q-Weighted model is positively shifted both for reward and no reward compared to classic Rescorla–Wagner when Q values are low. Conversely, for high values of Q, the change in Q in the Quadratic Q-Weighted model is negatively shifted both for reward and no reward compared to classic Rescorla–Wagner when Q values are high. Since the Rescorla–Wagner model asymptotically

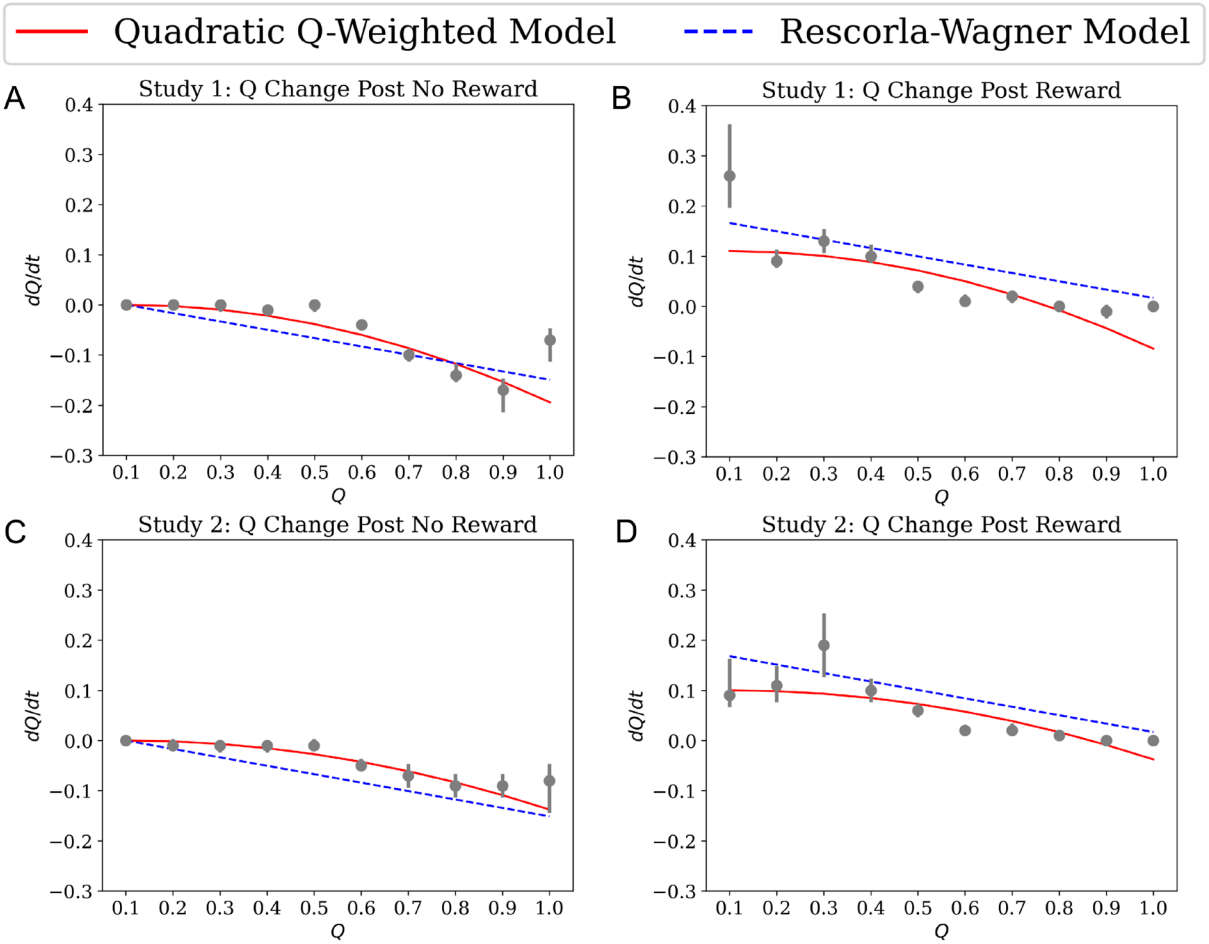

Fig. 2. An overview of behavior of the Quadratic Q-Weighted model we discovered using empirical data with SINDy. The x-axes reflect reported Q value and the y-axes are the median change in value. Gray dots show binned Q into 10 discrete categories, each with a bin size of 0.1. Categories were labeled with the upper bound of each bin. Error bars are 95% CI. (*A*) Study 1 empirical change in Q following no reward. (*B*) Study 1 empirical change in Q following reward. (*C*) Study 2 empirical change in Q following no reward. (*D*) Study 2 empirical change in Q following reward. Predicted changes in Q according to the best fit Quadratic Q-Weighted model (solid red) and the best fit Rescorla–Wagner model (dashed blue) are overlaid.

always converges to the true probability for any learning rate (48), the shifts shown in Fig. 2 demonstrate that the Quadratic Q-Weighted model implies the asymptotic bias in the estimation of the true probability. One example of the underestimation is that within this model an agent cannot predict Q values larger than the stable point $\sqrt{a/b}$ even when the reward probability is 1—this is where expectations stabilize. Noisy agents, like humans, can occasionally predict Q values larger than $\sqrt{a/b}$, but thereafter will be biased to lower their expectations back toward $\sqrt{a/b}$ even if met with further reward.

To further explore the implications of the Quadratic Q-Weighted model on participants' behavior, we employed linear mixed effects models to predict changes in expectations as a function of reward and distance from the stable point $\sqrt{a/b}$. We conducted a total of four models; two for each study, one of which included only post-reward trials and the other post-nonreward trials. The independent variable in each of the models was whether the current Q value was lower or higher than the stable point $\sqrt{a/b}$. The dependent variable was change in Q from previous trial. We dummy coded the model such that the data above the stable point would be the intercept of the model. This allowed us to not only compare significance between the conditions (above or below $\sqrt{a/b}$), but also compare results from above the stable point to zero. Our model also included a random variable of participant id. Starting with the intercept of the model, which compared the above the stable point results to zero, results suggested that in both Study 1 and Study 2, when receiving a reward and when they were above the stable point, participants significantly lowered their estimation of Q (Fig. 3 Orange bar compared to 0; Study 1: b = –0.167, $P < 0.001$; Study 2: b = –0.085, $P < 0.001$). These results would not have been seen if participants were using a classical Rescorla–Wagner model in which participants always increase their estimation of Q following a reward. Similar results were found in cases where there was no reward, such that when above the stable point, participants also significantly lowered their estimation of Q (Fig. 3 Orange bar compared to 0; Study 1: b = –0.291, $P < 0.001$; Study 2: b = –0.328, $P < 0.001$). These results should be expected, as both in our model and in a classic Rescorla–Wagner model, participants would lower their estimation of Q following a no-reward. Having established this difference from zero, results also suggested that in all cases, there was a significant difference in change in Q as a function of whether the previous Q was above or below the stable point (Study 1 Rewarded: b = 0.320, $P < 0.001$; Study 1 Unrewarded: b = 0.341, $P < 0.001$; Study 2 Rewarded: b = 0.223, $P < 0.001$; Study 2 Unrewarded: b = 0.337, $P < 0.001$). These results are congruent with Rescorla–Wagner.

Building on these findings, we next employed a complementary approach to balance the discovery of generalizable learning

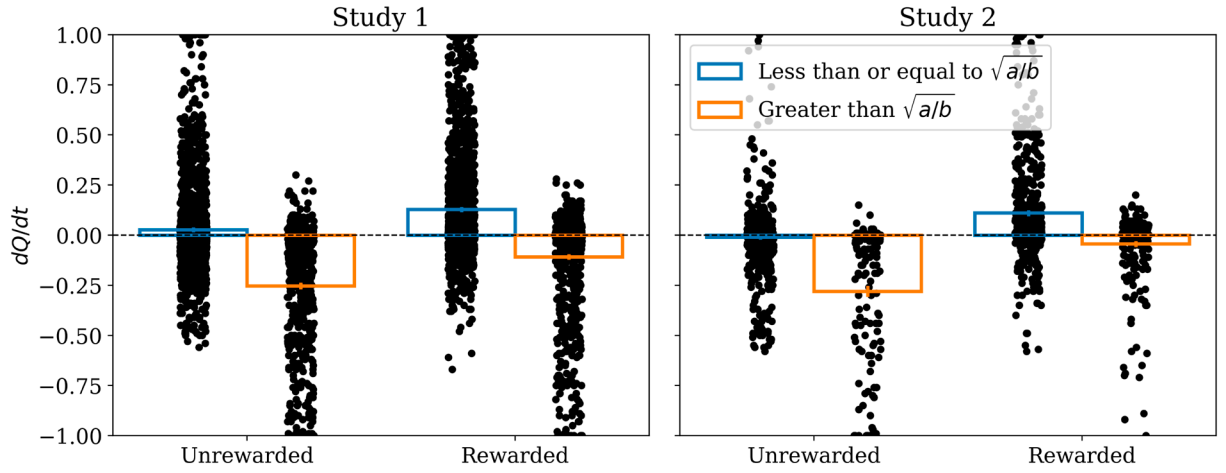

Fig. 3. Empirical changes in expectation *Q* as a function of *Q*'s position relative to the stable point ($\sqrt{a/b}$) and reward. Error bars are 95% CI. 10% of observations are included as dots to visualize the response distribution. Decreases in *Q* can be observed when *Q* is greater than the stable point, even following reward.

dynamics with the need to capture individual variability. Although the analysis conducted with the SINDy algorithm allowed us to identify the core functional form of a learning model by pooling data across participants, we recognize that pooling data in this way can obscure meaningful individual differences. To address this possibility, we next used the probabilistic programming language Stan (49) for individual-level model fitting, allowing us to assess the generalizability of the model and estimate participant-specific parameters. This also allowed us to validate the Quadratic Q-Weighted model's performance by comparing it against other existing models. Specifically, we fit five competing nonhierarchical models to each subject from our empirical data using Stan (see *SI Appendix* for model specifications and fitting procedures). These models included: a classic Rescorla–Wagner model (2), a Rescorla–Wagner model with time-dependent exponential decay (2), a Rescorla–Wagner model with asymmetric learning rates (41–44), a binary Kalman filter model (50), and SINDy's discovered model, the Quadratic Q-Weighted model. We chose to add a binary Kalman filter model to the analysis, despite the fact that it takes latent variables that cannot be discovered by SINDy, to get a sense of how the model compares to such modern models that include prediction uncertainty. We compared the relative fits of models using the Bayesian information criterion (BIC), which penalizes more complex models for the number of free parameters they include. The resulting BICs revealed that the Quadratic Q-Weighted model outperformed all alternative models (*SI Appendix*). A comparison of fits across models is visualized for a representative participant in Fig. 4. These results at the individual participant-level support our prior group-level analyses conducted with the support of SINDy: Participants' behavior in making probabilistic inference in our study is best explained by a SINDy discovered RL algorithm, the Quadratic Q-Weighted model.

Having established the Quadratic Q-Weighted model's superior fit at both the group and individual levels, we next examined the distribution of individualized parameters to explore the extent of heterogeneity in participants' learning behavior. Beyond the superior group-level fit indicated by BIC, we also find at the individual-level that the Quadratic Q-Weighted model best fit 68.35% of participants in Study 1, and 64.41% in Study 2. Importantly, the a and b coefficients scaling the reward and Q terms ($Q_{t+1} = ar_t - bQ_t^2$) were free to vary between participants in these fitted models. This revealed substantial heterogeneity among individuals around the group coefficient values discovered by SINDy: Study 1 mean a = 0.21 ± 0.18, Study 1 mean b = 0.45 ± 0.46; Study 2 mean a = 0.28 ± 0.19, Study 2 mean b = 0.61 ± 0.52. These findings suggest significant individual differences in how participants weigh recent rewards and adjust for previous estimates.

To further probe this variability, we allowed the exponent on Q to vary freely as an exploratory follow-up. This adjustment improved the model fit for 61.1% of participants in Study 1 and 77.96% in Study 2, indicating that the fixed exponent of 2 used in the original Quadratic Q-Weighted model does not fully capture all individual differences. These individualized exponents were estimated to be on average 1.52 ± 0.76 in Study 1 and 1.43 ± 0.81 in Study 2, indicating a high degree of variability among individuals in how they update expectations.

Phase 2: Evaluating Decision Models by assuming the Quadratic Q-Weighted Model in Existing Datasets. After finding the Quadratic Q-Weighted model with empirical data, we aimed to demonstrate its value by reanalyzing prior studies of choice behavior. Our goal was to determine whether the Quadratic Q-Weighted model provided a better account of learning in decision-making tasks than does the Rescorla–Wagner model. This reanalysis also provided an opportunity to evaluate the model's performance in studies that did not overtly measure participants' expectations—a vast majority of decision-making tasks do not probe estimates of probability overtly; they instead ask participants to act on implicit learned probabilities by selecting between two or more alternatives. Researchers are often most interested in the elements governing these selections, such as explore-exploit tendencies and stochasticity. However, because it is assumed that selection depends on one's estimates of reward probability, it is critical that researchers assume a learning model that best captures those estimates and their dynamics. To this end, we reanalyzed open datasets sourced from nine papers published in leading academic journals, each using the Rescorla–Wagner updating rule nested within larger decision models, with the goal of replacing that rule with our Quadratic Q-Weighted model. In each of the datasets, we used the authors' original analysis scripts for model fitting (see *SI Appendix* for details). To compare the authors' model and our variation with the Quadratic Q-Weighted model as a learning rule, we modified the original authors' scripts to fit a variation of their model using the general form of the Quadratic Q-Weighted model in place of Rescorla–Wagner. Notice that the model produced by SINDy in the empirical phase had a specific

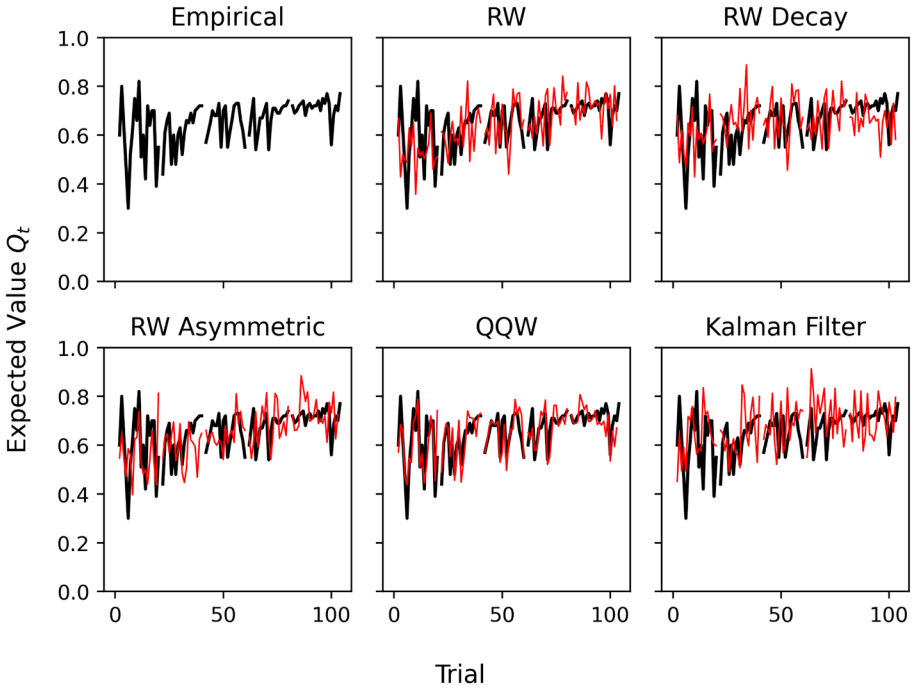

Fig. 4. Expected value estimates (Q_t) over 100 trials for a single representative participant. The "Empirical" panel represents the participant's reported values (black lines), while the remaining panels depict predictions (red lines) from models: RW (Rescorla–Wagner), RW with exponential decay, RW with asymmetric learning rates, QQW (Quadratic Q-Weighted), and the Kalman Filter. Missing data correspond to trials where attention checks were administered. Although all models demonstrate a generally good fit to the observed data, the QQW model stands out with a superior fit.

coefficient value for each term. In these analyses, we allowed those coefficients to vary freely between subjects.

To compare the model that was used in each paper to our Quadratic Q-Weighted model we calculated the BIC of both models using the summed likelihood estimates of each participant's data. BIC was chosen due to its consistency in identifying parsimonious yet well-fitting models by penalizing more heavily for superfluous parameterization. Likewise, BIC was used by all authors of the selected datasets. In all but one dataset, the model using the Quadratic Q-Weighted learning rule outperformed the original best model (Table 1).

Table 1. Model fits from each reanalyzed datasets using original authors' models and variations replacing Rescorla–Wagner learning rules with the Quadratic Q-Weighted model

	Original BIC	*Quadratic Q-Weighted BIC*
Kool et al. 2017 Experiment 1	461.35	**458.47**
Kool et al. 2017 Experiment 2	474.32	**450.23**
Lefebvre et al. 2017 Experiment 1	3857.55	**3806.44**
Lefebvre et al. 2017 Experiment 2	2512.77	**2496.62**
Palminteri et al. 2017 Experiment 1	3206.48	**3199.18**
Chambon et al. 2020 Experiment 4	**10987.38**	10997.49
Decker et al. 2016	37463.96	**36956.91**
Potter et al. 2017	25784.64	**25373.13**
Nussenbaum et al. 2020	63378.57	**62360.65**

Note: BIC is Bayesian Information Criterion. Lower BIC reflects better model fit. Original BIC is for the model used by the authors of the dataset. Quadratic Q-weighted BIC is for the variation of those models using the Quadratic Q-Weighted model as a nested learning rule in place of the Rescorla–Wagner learning rule.

Beyond fit, models using the Quadratic Q-Weighted learning rule may provide downstream benefits to understanding processes of decision-making. For example, one dataset in particular, Kool et al., 2017 Experiment 2 (51), the Quadratic Q-Weighted model yielded decision-making parameter estimates that diverged from the authors' main findings. In their 2017 article, Kool and colleagues (51) proposed that arbitration between model-based (MB) and model-free (MF) learning systems involves a cost–benefit analysis. The MB system, which plans toward goals, is more accurate but computationally demanding, whereas the MF system relies on habits and is computationally efficient but less flexible. Kool et al. suggested that people decide which system to use by weighing the benefits of the MB system's accuracy against its cognitive cost. This implies that MB control is employed when its benefits (in terms of reward) outweigh the costs (in terms of cognitive effort).

The most common method for measuring individual difference in this cost–benefit analysis is with the Two-Step Task (52), a complex, multistep decision-making task. Despite this, Kool et al. demonstrated in an earlier paper (53) that there exists no cost–benefit relationship between model-based vs. model-free strategy and performance on the task, and remediate this with a novel version of the task. To explore this further, they tested the effect of monetary stakes on strategy use in the original task, positing that high stakes should fail to yield increased MB control when it provides no advantage. They tested this by fitting a dual-systems RL model to their data (52). This RL model includes three separate Rescorla–Wagner updating rules for changing participants' expectations of reward followed by choosing specific spaceships and aliens. They found that there was no difference in MB control between high and low stakes on the original version of the task (Experiment 2; $t(99) = 0.4132$, $P = 0.6804$). In interpreting these findings, the authors suggest that participants might have a prior belief that MB control is generally associated with higher rewards, driven by real-world experiences where MB control is usually beneficial. This belief, reinforced by training, led participants to maintain a mix of MB and MF strategies.

We disagreed with this conclusion: if MB control is no better than MF on the original version of the task, participants should use MB control *less* than MF under high stakes. This is because high stakes can be stressful and impose cognitive load. MB control is costlier, despite it being equally as effective in this case, and the cognitive effort required to use MB control could be too demanding in high stakes situations. It is possible that parameter estimates could be improved and better reflect our hypothesized effect if a learning model were fit to the data which better reflects peoples' true learning. To test this, we modified the fitted dual systems RL model to instead use *the Quadratic Q-Weighted* model discovered by SINDy in our empirical studies. We found that our version of the dual systems model with the Quadratic Q-Weighted model as a learning rule better fit Kool et al.'s Experiment 2 data ($BIC_{QQW} = 450.23$) than the original dual systems RL model ($BIC_{RW} = 474.32$). Furthermore, we observed that the fitted free parameters for the MB weight in high stakes was indeed significantly less than the MB weight in low stakes ($t(99) = 2.9303$, $P = 0.0042$), supporting our hypothesis.

Discussion

In this work, we demonstrated that bottom–up equation discovery algorithms can be used for model development in social sciences. We collected empirical data in two variations of a learning task and used SINDy to develop an appropriate model for participants' behavior. This model—the Quadratic Q-Weighted model—provided insights into human learning and accounts for several interesting behavioral phenomena. Most importantly, the model introduces a tendency over the long term to underestimate Q values when true reward rates are high and overestimate Q values and true reward rates are low. Finally, we nested the Quadratic Q-Weighted model within existing, more complex decision models used by the authors of nine published datasets. These models notably were of decisions made on complex decision-making tasks, entirely different from our probability estimation task. We found that the revised models using our Quadratic Q-Weighted model instead of the Rescorla–Wagner updating rule provided a better fit in eight out of nine of those cases compared to the models originally used by the authors. We further demonstrate that the Quadratic Q-Weighted model does more than simply improve fit; it impacted the conclusions and interpretation of a previous study of complex decision-making. These results provide a promising path for the use of the Quadratic Q-Weighted model, as well as the use of equation discovery algorithms to development of interpretable but more predictive computational models.

Across multiple levels of analysis, what stands out as the most important feature of the Quadratic Q-Weighted model is its prediction that participants tend to overestimate low probabilities and underestimate high probabilities. This systematic bias may reflect the influence of persistent prior beliefs, consistent with Bayesian frameworks of inference (54, 55). In Bayesian models, a prior represents the learner's initial beliefs about the probability distribution of an outcome. If individuals maintain a strong prior centered around a moderate probability value (e.g., 0.5), subsequent learning will appear conservative, pulling extreme probabilities toward the center. This behavior is qualitatively similar to the effects produced by the quadratic term in our model, which attenuates updates as expectations approach the extremes of 0 and 1.

Recent work by Zhu et al. (56, 57) provides a compelling process-level account of such biases. Their Bayesian Sampler and Autocorrelated Bayesian Sampler models explain how cognitive constraints, such as limited sampling and autocorrelated internal states, can produce systematically biased probability estimates even when agents perform rational inference over time. These models show that conservatism and subadditivity can arise naturally from the sampling process, particularly when prior beliefs are strong and sampling is limited. While our model does not explicitly implement Bayesian sampling, it captures a similar behavioral signature: an asymptotic bias that emerges as participants anchor expectations toward intermediate values regardless of the observed evidence. This convergence raises interesting questions about whether the quadratic term in our model reflects a heuristic approximation of Bayesian updating or an internalized bias shaped by experience. Notably, we observe this pattern in both our empirical studies, including Study 2 where initial reward probabilities varied. Future work could directly compare the Quadratic Q-Weighted model to Bayesian sampling models, particularly under conditions that manipulate priors or restrict cognitive resources, to better understand the origins of such nonlinear updating dynamics.

While the current work focused on model development within RL, we envision many exciting new directions for model development in various social domains. Many subdomains in social sciences are still utilizing existing top–down models and these models can potentially be improved while maintaining interpretability. For example, models of social contagion, such the SIR epidemic models (58), which already benefit from further development using tools such as SINDy (24), can also be tested in explaining social interaction and contagion. Other domains such as decision making (59), planning (60), norm formation (61), affect (62), and many others all have models that are constantly being developed using top–down approaches, and these domains could potentially benefit from using equation discovery algorithms for model improvement and development. Finally, many domains in the social sciences involve analysis of longitudinal data, which is often analyzed using structural equation modeling or other tools that mostly test for linear processes (63). SINDy and other equation discovery tools are well suited to fit existing longitudinal data in order to uncover driving equations. It is important to note that we do not wish to eliminate hypothesis testing or theory-based models, but rather to expand the modeler's toolbox in considering alternative models for comparison and later confirmatory analyses, thus encouraging the discovery of novel, interpretable, predictive, and generalizable models.

Limitations and Future Directions

We acknowledge several limitations in using SINDy for model discovery in social science. First, our implementation is at present limited to modeling directly observable data. Across our empirical studies, Q-value was an explicit variable. In many RL experiments, expected value is a latent variable that is estimated from directly observable decisions (54, 64). Other models could include these latent variables, such as for uncertainty in expectations, which could potentially improve their quality in terms of predictability and generalizability. Despite these limitations, the approach we adopted here has the potential to change how models are developed in the social science.

A second limitation is that the SINDy algorithm is bounded by the specific decisions made in its implementation, such as the list of candidate functions and hyperparameters that govern the discovered model's sparsity (and hence control over its complexity). These decisions, as well as the indication of whether the discovered model is suitable, are subjective. Therefore, it is possible that there may exist other alternative models which could be better fitting. For example, the present study omitted candidate functions of continuous time. Although our discovered Quadratic Q-Weighted model provides a unique perspective on probability

weighting, it departs from the broader theoretical basis of RL models like Temporal Difference (TD) learning (2), which can be applied in real-time and have demonstrated strong links to learning processes in the brain (48, 65). The quadratic transform used here, although beneficial for capturing nonlinearity in probability estimates, does not provide the same foundation for understanding learning as a general, time-continuous process. However, it may be possible to incorporate the discovered nonlinearity within a TD learning framework. Specifically, future work could modify the TD update equations to include a transformed value function, such as f(V)=V + βV2, allowing us to retain the incremental, time-based learning properties of TD while introducing systematic biases that capture nonlinearities in human learning behavior. This would create a hybrid model that not only improves predictive performance but also preserves the dynamic learning structure of TD, potentially bridging these two perspectives effectively.

A third limitation of this study is that the model that was discovered by SINDy was based on a relatively narrow empirical task. Although the Quadratic Q-Weighted model showed excellent generalizability in predicting behavior across two new empirical studies and eight of nine reanalyzed datasets, its broader applicability requires further exploration. One notable exception to the Quadratic Q-Weighted model's superior performance was observed in a go/no-go dataset (66), where the Rescorla–Wagner model provided a better fit to the underlying learning process. We note that this dataset's structure—featuring a large number of trials per participant ($N = 600$) but a small number of participants ($N = 20$)—may have reduced its ability to robustly differentiate between models. However, the unique demands of go/no-go tasks, which rely heavily on inhibitory control, may inherently favor simpler models like Rescorla–Wagner. Unlike the other datasets analyzed, the go/no-go task may not highlight the nuanced reward–probability interactions or asymptotic behaviors that the Quadratic Q-Weighted model captures well. Moreover, the candidate features that we provided to SINDy for modeling learning in forced-choice tasks may not adequately capture learning dynamics unique to go/no-go tasks. These findings may suggest that tailoring candidate functions to task-specific dynamics is crucial for improving the generalizability and performance of discovered models.

Although it may be necessary to tailor models to task-specific dynamics, perhaps like those of the go/no-go task, it is equally as valuable that models generalize across a wider range of decision-making contexts. Importantly, we did show in our empirical studies that the identified Quadratic Q-Weighted model was robust to changes in initial reward probability and to varying levels of diffusion noise. However, the model's utility in tasks where reward probabilities remain static over time or where probabilities are more conservatively bounded has yet to be tested. Future studies could further evaluate the model under these conditions, such as the change-point detection studies conducted by Nassar and colleagues (67, 68), where participants are asked to predict continuously varying outcomes. These tasks involve different reward structures that could help determine whether the nonlinearity identified in our model captures more general aspects of human learning, particularly under conditions of uncertainty and dynamically changing environments. Future research should extend the model to such decision-making problems and explore additional nonlinearities, such as exponential or logarithmic transformations.

It is also important that future research consider edge cases. An important aspect of the Quadratic Q-Weighted model is its prediction of asymptotic behavior, particularly in situations where the model suggests an inherent bias in how individuals estimate extreme probabilities. Specifically, the model predicts that participants cannot estimate Q values beyond a certain stable point ($\sqrt{(a/b)}$), even in situations where the true reward probability is maximal (e.g., 1). This creates an opportunity for empirical validation, whereby edge cases can be specifically designed to test these predictions. For example, future experiments could involve a prolonged sequence of trials with a constant reward probability of 1 to determine whether participants' estimates truly converge at the stable point predicted by the model or if they adaptively reach the true value. If participants are found to be limited by this predicted asymptotic bias, it would lend further support to the model's validity. Conversely, if participants adapt beyond the predicted stable point, it may indicate the need for further refinement of the model. Such empirical tests would help to identify the conditions under which the model captures or fails to capture human learning behavior and highlight the importance of understanding model limitations within different RL environments. That said, we acknowledge that the model may fail to predict behavior under such rigid conditions. Although future studies have the potential to further validate the Quadratic Q-Weighted model, we suggest that the model's scope be interpreted with the constraints of the learning environment in mind.

Finally, a fourth limitation concerns the interpretation of nonlinearity in the Quadratic Q-Weighted model. Specifically, our formulation embeds nonlinearity directly in the updating rule—$Q_{t+1} = ar_t - bQ_t{}^2$—rather than in a utility transformation of outcomes prior to updating. While this structure captures state-dependent prediction errors, it differs computationally from models that apply a nonlinear utility function to feedback, such as $Q_{t+1} = au(r_t) - bQ_t$, where $u(r_t)$ might be quadratic. These two sources of nonlinearity reflect distinct psychological mechanisms: Utility curvature implies preferences over outcomes, while nonlinear updating implies that learning itself is biased or distorted based on prior expectations. Moreover, they make different predictions for learning. For example, nonlinear updating can produce attraction to stable points—values of Q where learning effectively ceases—whereas nonlinear utility does not predict such asymptotic behavior, as it preserves linear dependence on the current value estimate. That said, empirically distinguishing between these mechanisms is difficult within the current scope, as reward magnitudes in all of our studies (and those we reanalyzed) are discrete (either 0/1 or −1/1). Detecting utility curvature robustly would likely require tasks with continuous or graded rewards. While distortions akin to nonlinear utility (e.g., asymmetric weighting of 0 vs. 1 s) may partially be captured by the Quadratic Q-Weighted model's separate terms for reward and $Q_t{}^2$, this approach cannot model all plausible forms of utility transformation (e.g., inflation of 0 s). We therefore leave the question of disentangling nonlinear utility from nonlinear updating to future work. Promising directions include experimental designs using continuous outcomes and simulation-based model recovery studies that systematically vary utility and update structures in factorial combinations.

Overall, this study demonstrates the potential of bottom–up equation discovery methods, such as SINDy, to advance model development in the social sciences. The Quadratic Q-Weighted model provides key insights into human learning, uncovering systematic biases in probability estimation and generalizing across diverse datasets. By improving fit and influencing interpretations of prior studies, the model showcases the power of integrating nonlinear dynamics into decision-making frameworks. More broadly, this work highlights how data-driven discovery can complement theory-driven approaches, offering a path toward more interpretable and predictive models that deepen our understanding of human behavior.

Methods

Explaining SINDy. For our analysis, we used PySINDy (69, 70), a Python package that provides tools for applying the SINDy algorithm (71) for model discovery. SINDy is used to approximate the unknown governing equations of a dynamical system using a sparse regression framework. It assumes that the dynamics can be expressed as a sum of known functions, multiplied by unknown coefficients. By leveraging sparsity-promoting techniques, SINDy aims to identify the most relevant terms in the equation, effectively providing a parsimonious representation of the system's behavior. Typically, SINDy estimates derivatives for system variables of the following form:

$$\frac{d\boldsymbol{X}}{dt} = \theta(\boldsymbol{X}, \boldsymbol{U})\boldsymbol{B},$$

where $\boldsymbol{X}$ is a matrix of observed variables to be modeled, $\boldsymbol{U}$ is a matrix of control variables that are not to be modeled but may be important for modeling variables in $\boldsymbol{X}$, $\theta(X, U)$ is a matrix of candidate features selected by the researcher that transform the data, and $\boldsymbol{B}$ is a vector of coefficients that scale the candidate features. Through sparsification techniques such as L2 ridge regression, most of these coefficients are reduced to zero and only the most predictive features remain (see below for details on how we chose to promote sparsity). The objective of the SINDy algorithm is to solve for $\boldsymbol{B}$ given the researcher selected candidate features and the approximated first-Order derivatives of the observed data.

For most experiments in the social sciences, observed data are collected in discrete trials. Therefore, we used SINDy to estimate discrete-time models of the form:

$$x_{k+1} = F(x_k, u_k),$$

where x_k are the observed variables to be modeled from $\boldsymbol{X}$ at timepoint k, and u_k are the control variables (such as r and t in this case) at timepoint k. Rather than calculating a system of derivatives, using SINDy we calculated a matrix $\boldsymbol{X}'$ where the columns of $\boldsymbol{X}'$ are measures of x moved forward in time until the final datapoint at time K (i.e., $[\dot{x}_1, \dot{x}_2, \dot{x}_3, \cdots, \dot{x}_K]$). With this approach, SINDy estimates discrete-time equations for system variables in the form:

$$\boldsymbol{X}' = \theta(\boldsymbol{X}, \boldsymbol{U})\boldsymbol{B}.$$

We solve for $\mathbf{B}$ by using a variation of stepwise sparse regression (SSR) (72) to minimize the objective function:

$$\| \boldsymbol{X}'_k - \theta(\boldsymbol{X_k}, \boldsymbol{U_k}) b_k \|_2^2 + \lambda \| b_k \|_2^2,$$

where each element of the coefficient vector B is regularized with the L2 ridge regression value λ in the penalty term. We chose $\lambda = 0.2$ for all analyses (simulations and empirical data). Selecting the value of λ is crucial as it determines the trade-off between accuracy and parsimony; it is a hyperparameter that should be tuned by the modeler. On each iteration of the minimization procedure, SINDy first solves a standard least square regression to obtain a tentative, nonsparse solution $\mathbf{b}$:

$$\boldsymbol{b} = \underset{b_k \in \mathbb{R}^K}{\operatorname{argmin}} \| \boldsymbol{X}'_k - \theta(\boldsymbol{X_k}, \boldsymbol{U_k}) b_k \|_2^2.$$

All possible $\mathbf{b}$ are considered, each with one coefficient set to zero, and the solution $\mathbf{b}$ with the smallest residual error is selected. Least-squares regression is then again performed on the remaining degrees of freedom. This process continues until there is only one coefficient remaining. Next, we iterate in reverse order over the history of solutions starting from the simplest solution where all but one coefficient is set to 0. We continue to add non-0 coefficients back to the model until the next change in residual error is less than 0.05 times the previous iteration's residual error, at which point the process terminates, providing a sparse solution vector $\mathbf{B}$. This provides a solution that is most parsimonious with the least loss. Like λ, this multiplier of 0.05 is also a hyperparameter and can be tuned to select a sparsity level for the solution that neither under- or overfits the data.

Simulations. The first step of any SINDy analysis is collecting or simulating data to populate the time-series observational data $\mathbf{X}$ and $\mathbf{U}$. To investigate the applicability of SINDy in recovering the governing equations of established RL models, we generated synthetic data through simulations. We selected well-known RL models–the Rescorla–Wagner model and variants–as the basis for generating the data. See *SI Appendix*, Equation Recovery from Simulated Data for details. Our simulated agents estimated the probability of reward as $\mathbf{X}$. The history of reward and trial number were represented as separate columns in $\mathbf{U}$. From these observations, we used SINDy to calculate $\mathbf{X}'$ as the matrix of discrete-time variables x_k shifted x_{k+1}. To ensure that SINDy's solution $\mathbf{B}$ was interpretable, we provided SINDy with the following matrix of candidate functions:

$$\boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{U}) = \left[Q, r, t, Q^2, t^2, Q\times r, Q\times t, r\times t, \frac{1}{t+100}, e^{-\frac{t}{30}}, e^{-\frac{t}{20}} \cdots\right]$$

$$\left[\cdots e^{-\frac{t}{10}}, \frac{Q}{t+100}, Q\times e^{-\frac{t}{30}}, Q\times e^{-\frac{t}{20}}, Q\times e^{-\frac{t}{10}}, \frac{r}{t+100}, |r-Q|, \left|r-Q^2\right| \cdots\right]$$

$$\left[\cdots r\times e^{-\frac{t}{30}}, r\times e^{-\frac{t}{20}}, r\times e^{-\frac{t}{10}}, \frac{t}{t+100}, t\times e^{-\frac{t}{30}}, t\times\times e^{-\frac{t}{20}}, t\times e^{-\frac{t}{10}} \cdots\right].$$

These functions were chosen since they provide necessary components to allow SINDy to discover existing theories of learning. Of course, SINDy can also find novel combinations of these candidate functions. *Q and r* are the building blocks for the model and serve as basic variables for delta-updating rules (2) and as changes may vary over time we also added *t* as a potential variable. A few decay rates (e.g., −t/10, −t/20) were chosen to allow for a variety of exponential shapes. We initialized the model with linear terms to allow for a classic Rescorla–Wagner model. We also introduced quadratic terms as we wanted to test whether changes in prediction of *Q* or in the evaluation of *r* may not be linear, in line with the notion that stimuli is represented by individuals with some form of power transformation (73). We also wanted to examine whether the two terms, *Q* and *r,* interact with each other or with time, and whether the absolute difference between *Q* and *r* might modulate learning. Finally, we wanted to allow changes in both *Q and r to* decay at different rates. As can be seen from this list, we were conservative in our function choices in order to make sure that the resulting model was interpretable. Future work may use other terms in line with the researcher's goals and evaluation of the appropriate relationship between variables.

Phase 1: Equation Discovery from Empirical Probability Estimates.

Participants. All experiments received IRB approval from Harvard Business School (IRB22-0546) and informed consent was obtained from all research participants. In setting our sample size for Study 1, we decided to start with a large sample of 500 participants in order to determine the appropriate sample required for SINDy to make accurate predictions. We recruited participants through Prolific. Participants were paid $4 for their participation in the study in addition to $.03 for every time that their estimate was within 5% of the true reward rate. Participants were given attention checks to ensure data quality. On the first and fiftieth trials, participants were asked to type a predetermined word. On the second trial, and every twenty trials thereafter, participants were asked to respond to the slider scale with a specific percentage. On these attention check trials, participants were instead prompted with "This is an attention check. Please move the slider to _%," where the percentage was a number between 0 and 100. Participants who failed either of the word typing checks or more than 2 of the slider checks were excluded from analysis. Because we assumed that some participants would not be able to complete the task, we recruited a larger number of participants than required N = 543. Our exclusion criteria were conditioned on attention checks (*Attention Checks and Exclusion Criteria*). We removed three participants for failed word typing checks and 85 for failed slider checks, for a final sample of $N = 455$ (men: 216, women: 216, other or refused to answer: 23; age, $M = 36.25$, $SD = 12.94$).

After establishing the results from Study 1, we conducted an analysis to test the appropriate sample size required for SINDy to capture the appropriate model. We randomly sampled participants from Study 1 to find the smallest *N* necessary to reliably recover the Quadratic Q-weighted model. 100 iterations of this sampling procedure suggested that ~200 participants were enough for SINDy to reliably capture the model. In Study 2, we therefore aimed for N = 200. We again used Prolific for recruitment and paid participants the same sum as in Study 1. Our initial sample was 206. We used the same selection criteria for exclusions. We removed 29 participants for failed slider checks, for a final sample was $N = 177$ (Men: 87, Women: 85, other or refused to answer: 5; Age, $M = 37.88$, $SD = 12.06$).

Task. We used jsPsych (74) to conduct our study. Participants logged in and were told to imagine that they are inspecting a factory that produces phones. The factory produces phones one at a time and will then be inspected by the participant. Participants were told that the phone would either be working or defective. Participants were told that they are asked to estimate the probability that the next phone will be a working phone (Fig. 2). After a single practice trial, participants completed 100 trials of the task. Participants were first presented with an inspection slide, depicting a factory and a phone with a "?" printed on it. Participants were to click a button labeled "Inspect" below the images without any time constraints. After clicking the Inspect button the phone with the ? printed on it was revealed to either be working with a green check mark, or defective with a red "X." Participants were required to observe this feedback screen for 3 s before they could advance the page. On the following slide, participants were asked to respond on a slider scale from 0 to 100% what they believed the probability was that the next phone would be a working phone. Participants had to move the slider from its initial value (50%) in order to make their prediction. Participants were given as much time as they needed to make this judgment, but were required to wait 3 s before they could respond. This completed a trial and was repeated for a total 100 trials. A fixation cross was presented for 1 s between trials. After completing the task, participants were sent to a Qualtrics survey to fill in their demographics.

Measures. When participants completed the learning task, they were asked to estimate the probability that the next phone will be defected. After completing the task, participants filled out a TIPI 10-item personality measure (75) and a short demographics survey in which they were asked for their name, age, gender, race, ethnicity, first language, political affiliation, citizenship, nation of birth, annual income, and email address. See *SI Appendix* for full analysis of demographics.

Phase 2: Evaluating Decision Models by Assuming the Quadratic Q-Weighted Model in Existing Datasets.

Paper selection. We identified papers and datasets for reanalysis using the Niv Lab OpenData repository (https://nivlab.github.io/opendata/). Tags "2-arm bandit," "restless bandit," and "two-step" were considered. Criteria for reanalysis included 1) having freely accessible trial-level data, 2) having freely accessible code for model fitting and analysis, 3) the use of a Rescorla–Wagner delta-updating rule nested within the fitted model, and 4) association with a published paper. Many datasets did not meet the criteria, narrowing our search to the following nine datasets:

Kool et al., 2017 Experiments 1 & 2 (51). Published in Psychological Science (https://osf.io/yg82m/). The goal of this project was to examine whether people choose between model-free versus model-based control based on a cost–benefit analysis. The task was based on the Daw two-step decision making task (52). Participants made a first choice between two spaceships (green or blue), each leading to two planets with different probabilities in each of the studies (red, or purple). In Study 1, the probability of getting to a certain planet with a certain ship was always 100%. In Study 2 the probability of getting to one of the two planets was always 70% for each spaceship. When they arrived at the planet, participants met either one alien (Experiment 1) or chose one of two aliens (Experiment 2) who gave them a reward. Aliens were either in a good or a bad part of a mine and the probability of quality of their reward changed over time (drift rate in reward). In both studies, the researchers manipulated the size of the reward (stakes: 1 point or 5 points).

Lefebvre et al., 2017 Experiments 1 & 2 (76). Published in Nature Human Behavior; *Palminteri* et al., *2017 Experiment 1* (77). Published in PLOS Computational Biology; and *Chambon et al., 2020 Experiment 4* (66). Published in Nature Human Behavior. Each of these datasets and analysis code were acquired from a meta-analytic study, Palminteri, 2023 (78) (https://github.com/spalminteri/conf-bias-meta-analysis). We followed the models of Palminteri, 2023 for reanalyzing each of these four datasets. Shared among the authors was an interest in the confirmation bias hypothesis, in which a person learns more from positive reinforcement that supports their preexisting biases than they do negative reinforcement disproving their beliefs. The tasks were all variants of a simple two-armed bandit. Participants chose between two alternatives presented as symbols and either received or did not receive reward. Each bandit had a predesignated probability of reward but in all tasks, participants observed the outcome of their chosen symbol, but received no information from the unchosen symbol. Lefebvre et al., 2017 Experiments 1 & 2 used probabilities 25%/75%, Chambon et al., 2020 Experiment 4 used 30%/70%, and Palminteri et al. Experiment 1 used 50%/50%, 25%/75%, and a 17%/83%. These probabilities reversed halfway through the task, depending on assignment to experimental conditions.

Decker et al., 2016 (79). Published in Psychological Science; *Potter et al., 2017* (80). Published in Developmental Cognitive Neuroscience; and *Nussenbaum* et al., *2020* (81) Published in Collabra: Psychology. Each of these three datasets and analysis code were acquired from a reanalysis conducted by Nussenbaum et al., 2020 (https://osf.io/we89v/). The authors all investigated the emergence of model-based control across development, using the classic version of the Two-Step task as a propensity measure of model-based control. The task was identical to the version used in Kool et al., 2017 Experiment 2, without manipulating the size of reward.

Data, Materials, and Software Availability. All simulation and empirical data are available on the Open Science Framework here: https://osf.io/aeujf/?view_only=88b2b75499f54a3895502fc353f4d244 (82). All analysis scripts and modeling code are available on GitHub here: https://github.com/GoldenbergLab/analysis-rl-sindy-kyle (83).

Author affiliations: [a]Department of Psychological Sciences, Case Western Reserve University, Cleveland, OH 44106; [b]Booth School of Business, University of Chicago, Chicago, IL 60637; [c]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]Department of Psychology, Harvard University, Cambridge, MA 02138; [e]Graduate School of Business, Stanford University, Stanford CA 94305; [f]Harvard Business School, Harvard University, Boston, MA 02163; and [g]Digital, Data and Design Institute, Harvard University, Cambridge, MA 02138

1. O. Guest, A. E. Martin, How computational modeling can force theory building in psychological science. *Perspect. Psychol. Sci.* **16**, 789–802 (2021).
2. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998).
3. T. Akam, M. E. Walton, What is dopamine doing in model-based reinforcement learning? *Curr. Opin. Behav. Sci.* **38**, 74–82 (2021).
4. B. B. Doll, D. A. Simon, N. D. Daw, The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
5. S. J. Gershman, N. Uchida, Believing in dopamine. *Nat. Rev. Neurosci.* **20**, 703–714 (2019).
6. E. O. Neftci, B. B. Averbeck, Reinforcement learning in artificial and biological systems. *Nat. Mach. Intell.* **1**, 133–143 (2019).
7. Y. Niv, Reinforcement learning in the brain. *J. Math. Psychol.* **53**, 139–154 (2009).
8. M. Campbell, A. J. Hoane, F. Hsu, Deep Blue. *Artif. Intell.* **134**, 57–83 (2002).
9. D. Silver *et al.*, Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
10. O. Vinyals *et al.*, Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
11. J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, 1944).
12. R. D. Luce, On the possible psychophysical laws. *Psychol. Rev.* **66**, 81–95 (1959).
13. A. Tversky, D. Kahneman, Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–291 (1979).
14. D. Prelec, The probability weighting function. *Econometrica* **66**, 497–527 (1998).
15. R. Gonzalez, G. Wu, On the shape of the probability weighting function. *Cogn. Psychol.* **38**, 129–166 (1999).
16. J. C. Denrell, Reference-dependent risk sensitivity as rational inference. *Psychol. Rev.* **122**, 461–484 (2015).
17. D. U. Wulff, M. Mergenthaler-Canseco, R. Hertwig, A meta-analytic review of two modes of learning and the description-experience gap. *Psychol. Bull.* **144**, 140–176 (2018).
18. A. Glöckner, B. E. Hilbig, F. Henninger, S. Fiedler, The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *J. Exp. Psychol. Gen.* **145**, 486–508 (2016).
19. R. A. Rescorla, A. R. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement" in *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokasy, Eds. (Appleton-Century-Crofts, 1972).
20. Y. Niv, J. A. Edlund, P. Dayan, J. P. O'Doherty, Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32**, 551–562 (2012).
21. C. M. Constantinople, A. T. Piet, C. D. Brody, An analysis of decision under risk in rats. *Curr. Biol.* **29**, 2066–2074.e5 (2019).
22. A. Tymula *et al.*, Dynamic prospect theory: Two core decision theories coexist in the gambling behavior of monkeys and humans. *Sci. Adv.* **9**, eade7972 (2023).
23. M. K. Eckstein, L. Wilbrecht, A. G. Collins, What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Curr. Opin. Behav. Sci.* **41**, 128–137 (2021).
24. J. Horrocks, C. T. Bauch, Algorithmic discovery of dynamic models from infectious disease data. *Sci. Rep.* **10**, 7061 (2020).

25. L. K. Bartlett, A. Pirrone, N. Javed, F. Gobet, Computational scientific discovery in psychology. *Perspect. Psychol. Sci.* **18**, 178–189 (2023).
26. A. Almaatouq *et al.*, Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* **47**, 1–55 (2022), 10.1017/S0140525X22002874.
27. M. Fintz, M. Osadchy, U. Hertz, Using deep learning to predict human decisions and using cognitive models to explain deep learning models. *Sci. Rep.* **12**, 4736 (2022).
28. J. M. Hofman *et al.*, Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
29. L. Ji-An, M. K. Benna, M. G. Mattar, Automatic discovery of cognitive strategies with tiny recurrent neural networks. bioRxiv [Preprint] (2023). https://www.biorxiv.org/content/10.1101/2023.04.12.536629v2 (Accessed 27 July 2023).
30. P. I. Jaffe, R. A. Poldrack, R. J. Schafer, P. G. Bissett, Modelling human behavior in cognitive tasks with latent dynamical systems. *Nat. Hum. Behav.* **7**, 986–1000 (2023).
31. N. A. Roy, J. H. Bak, A. Akrami, C. D. Brody, J. W. Pillow, Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron* **109**, 597–610.e6 (2021).
32. K. J. Miller, M. Eckstein, M. M. Botvinick, Z. Kurth-Nelson, Cognitive model discovery via disentangled RNNs. bioRxiv [Preprint] (2023). https://www.biorxiv.org/content/10.1101/2023.06.23.546250v1 (Accessed 16 July 2023).
33. M. Agrawal, J. C. Peterson, T. L. Griffiths, Scaling up psychology via scientific regret minimization. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8825–8835 (2020).
34. J. Bongard, H. Lipson, Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9943–9948 (2007).
35. M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
36. E. Kaiser, J. N. Kutz, S. L. Brunton, Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proc. R. Soc. A., Math., Phys. Eng. Sci.* **474**, 20180335 (2018).
37. N. M. Mangan, J. N. Kutz, S. L. Brunton, J. L. Proctor, Model selection for dynamical systems via sparse regression and information criteria. *Proc. R. Soc. A., Math., Phys. Eng. Sci.* 473, 20170009 (2017).
38. S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
39. E. P. Alves, F. Fiuza, Data-driven discovery of reduced plasma physics models from fully kinetic simulations. *Phys. Rev. Res.* **4**, 033192 (2022).
40. K. Kaheman, E. Kaiser, B. Strom, J. N. Kutz, S. L. Brunton, Learning discrepancy models from experimental data. arXiv [Preprint] (2019). https://arxiv.org/abs/1909.08574 (Accessed 18 April 2023).
41. Z. Lai, S. Nagarajaiah, Sparse structural system identification method for nonlinear dynamic systems with hysteresis/inelastic behavior. *Mech. Syst. Signal Process.* **117**, 813–842 (2019).
42. M. Sorokina, S. Sygletos, S. Turitsyn, Sparse identification for nonlinear optical communication systems: Sino method. *Opt. Express* **24**, 30433–30443 (2016).
43. N. M. Mangan, S. L. Brunton, J. L. Proctor, J. N. Kutz, Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2**, 52–63 (2016).
44. R. Dale, H. S. Bhat, Equations of mind: Data science for inferring nonlinear dynamics of socio-cognitive systems. *Cogn. Syst. Res.* **52**, 275–290 (2018).
45. A. M. Bornstein, M. W. Khaw, D. Shohamy, N. D. Daw, Reminders of past choices bias decisions for reward in humans. *Nat. Commun.* **8**, 15958 (2017).
46. N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, R. J. Dolan, Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
47. M. Speekenbrink, E. Konstantinidis, Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).
48. J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, R. J. Dolan, Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).
49. B. Carpenter *et al.*, Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
50. P. Piray, N. D. Daw, A simple model for learning in volatile environments. *PLoS Comput. Biol.* **16**, e1007963 (2020).
51. W. Kool, S. J. Gershman, F. A. Cushman, Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol. Sci.* **28**, 1321–1333 (2017).
52. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
53. W. Kool, F. A. Cushman, S. J. Gershman, When does model-based control pay off?. *PLoS Comput. Biol.* **12**, e1005090 (2016).
54. A. C. Courville, N. D. Daw, D. S. Touretzky, Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* **10**, 294–300 (2006).
55. F. Meyniel, S. Dehaene, Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E3859–E3868 (2017).
56. J.-Q. Zhu, A. N. Sanborn, N. Chater, The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychol. Rev.* **127**, 719–748 (2020).
57. J.-Q. Zhu, J. Sundh, J. Spicer, N. Chater, A. N. Sanborn, The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychol. Rev.* **131**, 456–493 (2024).
58. O. W. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A, Math., Phys. Eng. Sci.* **115**, 700–721 (1927).
59. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
60. M. K. Ho *et al.*, People construct simplified mental representations to plan. *Nature* **606**, 129–136 (2022).
61. R. X. D. Hawkins, N. D. Goodman, R. L. Goldstone, The emergence of social norms and conventions. *Trends Cogn. Sci.* **23**, 158–169 (2019).
62. J. Gratch, S. Marsella, "Tears and fears: Modeling emotions and emotional behaviors in synthetic agents" in *Proceedings of the Fifth International Conference on Autonomous Agents* (2001), pp. 278–285, https://doi.org/10.1145/375735.376309.
63. J. Ullman, Structural equation modeling: Reviewing the basics and moving forward. *J. Pers. Assess.* **87**, 35–50 (2006).
64. S. J. Gershman, A unifying probabilistic view of associative learning. *PLoS Comput. Biol.* **11**, 1–20 (2015).
65. B. Seymour *et al.*, Temporal difference models describe higher-order learning in humans. *Nature* **429**, 664–667 (2004).
66. V. Chambon *et al.*, Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nat. Hum. Behav.* **4**, 1067–1079 (2020).
67. M. R. Nassar, R. C. Wilson, B. Heasly, J. I. Gold, An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* **30**, 12366–12378 (2010).
68. M. R. Nassar *et al.*, Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.* **15**, 1040–1046 (2012).
69. B. M. de Silva *et al.*, PySINDy: A Python package for the sparse identification of nonlinear dynamics from data. arXiv [Preprint] (2020). https://arxiv.org/abs/2004.08424 (Accessed 27 July 2023).
70. A. A. Kaptanoglu *et al.*, PySINDy: A comprehensive Python package for robust sparse system identification. *JOSS* **7**, 3994 (2022).
71. S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3932–3937 (2016).
72. L. Boninsegna, F. Nüske, C. Clementi, Sparse learning of stochastic dynamical equations. *J. Chem. Phys.* **148**, 241723 (2018).
73. S. S. Stevens, On the psychophysical law. *Psychol. Rev.* **64**, 153–181 (1957).
74. J. R. de Leeuw, R. A. Gilbert, B. Luchterhandt, JsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *J. Open Source Softw.* **8**, 5351 (2023).
75. S. D. Gosling, P. J. Rentfrow, W. B. Swann, A very brief measure of the big-five personality domains. *J. Res. Pers.* **37**, 504–528 (2003).
76. G. Lefebvre, M. Lebreton, F. Meyniel, S. Bourgeois-Gironde, S. Palminteri, Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1**, 1–9 (2017).
77. S. Palminteri, G. Lefebvre, E. J. Kilford, S.-J. Blakemore, Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Comput. Biol.* **13**, e1005684 (2017).
78. S. Palminteri, Choice-confirmation bias and gradual perseveration in human reinforcement learning. *Behav. Neurosci.* **137**, 78–88 (2023).
79. J. H. Decker, A. R. Otto, N. D. Daw, C. A. Hartley, From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychol. Sci.* **27**, 848–858 (2016).
80. T. C. S. Potter, N. V. Bryce, C. A. Hartley, Cognitive components underpinning the development of model-based learning. *Dev. Cogn. Neurosci.* **25**, 272–280 (2017).
81. K. Nussenbaum, M. Scheuplein, C. V. Phaneuf, M. D. Evans, C. A. Hartley, Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. *Psychol. Collabra* **6**, 17213 (2020).
82. K. J. LaFollette, J. Yuval, R. Schurr, D. Melnikoff, A. Goldenberg, Data from "Data-driven equation discovery reveals nonlinear reinforcement learning in humans." Open Science Framework. https://osf.io/aeujf/. Deposited 12 November 2023.
83. K. J. LaFollette, J. Yuval, R. Schurr, D. Melnikoff, A. Goldenberg, Analysis code from "Data-driven equation discovery reveals nonlinear reinforcement learning in humans." GitHub. https://github.com/GoldenbergLab/analysis-rl-sindy-kyle. Deposited 24 April 2025.