Predictive and Interpretable: Combining Artificial Neural Networks and Classic Cognitive Models to Understand Human Learning and Decision Making

Maria K. Eckstein[1] **(mariaeckstein@deepmind.com)**

Christopher Summerfield[1] **(chris@deepmind.com)**

Nathaniel D. Daw[1] **(ndaw@deepmind.com)**

Kevin J. Miller[1]2 **(ndaw@deepmind.com)**

[1]Natural Intelligence Team, Google DeepMind, London, UK [2]University College London, Institute of Ophthalmology, London, UK

Abstract

Quantitative models of behavior are a fundamental tool in cognitive science. Typically, models are hand-crafted to implement specific cognitive mechanisms. Such "classic" models are interpretable by design, but may provide poor fit to experimental data. Artificial neural networks (ANNs), on the contrary, can fit arbitrary datasets at the cost of opaque mechanisms. Here, we adopt a hybrid approach, combining the predictive power of ANNs with the interpretability of classic models. We apply this approach to Reinforcement Learning (RL), beginning with classic RL models and replacing their components one-by-one with ANNs. We find that hybrid models can provide similar fit to fully-general ANNs, while retaining the interpretability of classic cognitive models: They reveal reward-based learning mechanisms in humans that are strikingly similar to classic RL. They also reveal mechanisms not contained in classic models, including separate reward-blind mechanisms, and the specific memory contents relevant to reward-based and reward-blind mechanisms.

Keywords: computational cognitive modeling; recurrent neural networks; reinforcement learning; model comparison; model inspection

Introduction and Background

Computational models of behavior are a fundamental tool in many areas of cognitive science. Some of the most popular models include reinforcement learning (RL; Daw, 2011; R. R. Miller et al., 1995; Sutton and Barto, 2017), Bayesian inference (Ma, 2019), and evidence accumulation (Ratcliff and McKoon, 2008). Computational models have a long and rich history in the cognitive sciences, partly because they hold the tantalizing promise to quantitatively test elaborate hypotheses about unobservable cognitive processes.

To this aim, computational models typically embody hand-crafted hypotheses about a cognitive process of interest (Busemeyer and Diederich, 2010). For example, RL models assume that long-term reward histories are compressed into a low-dimensional, Markovian value representation, which is incrementally updated after new reward experiences, following a delta rule (Daw, 2011; Sutton and Barto, 2017). However, such classical models face problems (Eckstein et al., 2021):

1) It is fundamentally unclear *how well* any given model fits the data (Box, 1979; Navarro, 2019). Using the classic approach, several competing models are usually compared in terms of fit to identify the best (Katahira, 2016; Wilson and Collins, 2019). However, it is unclear when to stop searching: How do we determine that a winning model fits the data "well enough" (Palminteri et al., 2017)?

2) There also is the danger of model misspecification, e.g., of overlooking a mechanism that is crucial in the true data generating process (Nassar and Frank, 2016; Nussenbaum and Hartley, 2019). If a model is misspecified, we are not only disregarding a crucial mechanism, but other, correctly specified mechanisms will compensate for the deficit, leading to additional distortions (Sugawara and Katahira, 2021).

In this study, we argue that deep learning, amongst many possible applications in the cognitive sciences (Barak, 2017; Ma and Peters, 2020), can address these issues: 1) The universal function approximation theorem states that any sufficiently powerful artificial neural network (ANN) can approximate any input-output function arbitrarily well (Cybenko, 1989; Hornik, 1991): Sufficiently deep ANNs, trained on sufficiently large datasets, should therefore be able to represent any computational processes (LeCun et al., 2015), including cognitive computational processes. Recent studies have put this claim to test, and indeed found near-optimal fits of ANNs in simulation (Ger et al., 2023). In empirical data, many studies have shown extensive improvements in fit compared to hand-crafted models (e.g., Agrawal et al., 2020; Battleday et al., 2020; Kuperwajs et al., 2022; Peterson et al., 2021; Sutskever and Nair, 2008), including in RL tasks (Dezfouli, Ashtiani, et al., 2019; Dezfouli, Griffiths, et al., 2019; Fintz et al., 2021; Jaffe et al., 2022; Song et al., 2017; Yang et al., 2019).

2) ANNs offer the promise of overcoming model misspecification, precisely because they do not put constraints on the underlying mechanism, and hence allow any mechanism to be discovered. However, this flexibility also is a drawback: ANNs are often called "black boxes" because they do not—unlike hand-crafted models—offer direct insight into the process they learn to emulate (Ma and Peters, 2020).

Here, we first fit a classic hand-crafted RL model and an ANN to the same human learning task (Bahrami and Navajas, 2020). We formulate the RL model as a special case of the general ANN, which allows us to construct "hybrid" RL-ANN models to fill the gaps between both extremes: We will endow the RL model—step-by-step—with specific, interpretable, and delineated mechanisms until we obtain the fully general ANN. In this process, we calculate model fits to assess the empirical evidence for each added mechanism

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

in human behavior. We also probe and inspect the component mechanisms of each fitted model: For example, we can directly compare an ANN's learned input-output mappings to the fixed input-output mappings of the classic RL model. The main contribution of this study is to create cognitive models that capture human learning exhaustively, yet provide insight into its mechanisms.

Results

Dataset

Human Task In a publicly available dataset, 965 human participants completed 150 trials of a drifting 4-armed bandit task (Bahrami and Navajas, 2020; original task by Daw et al., 2006). On each trial, participants were asked to select one of four bandits and received a continuous numerical reward (1-98 points) based on the chosen bandit's current reward payout. Reward payouts drifted over time (Gaussian walk). Each participant was randomly assigned to one of three predetermined reward payoff schedules. For details, refer to Bahrami and Navajas (2020) and Daw et al. (2006).

Previous Findings The task was originally designed to study the neural mechanisms underlying exploratory and exploitative choice (Daw et al., 2006). More recently, Fintz et al. (2021) trained ANN models that suggest that participants used RL-like, reward-sensitive as well as exploration-like, reward-insensitive strategies. However, the study did not address how both mechanisms interact, a question we will address here.

Data Preprocessing We removed all participants who had missed more than 10% of trials (57 participants, 5.9%), a visually-determined elbow point, for a final sample size of 908. We set aside 20 participants per payoff schedule (60 total) as a testing dataset to assess model fit. Of the remaining 848 participants, we selected the largest balanced sub-sample in terms of reward schedules (825 participants; 275 per schedule) for the training dataset. Reward magnitudes were divided by 100 (resulting range: 0.01-0.98).

Modeling

Modeling Strategy In its basic form, the classic RL model $RL_{\alpha\beta}$ comprises two functions: A classic, fixed value update rule f_{RL} that calculates action values $v(a)$ based on observed rewards r_t; and a softmax action rule that translates values $v(a)$ into action probabilities $p(a)$ (Table 1). Our first hybrid model replaces the classic, fixed value update f_{RL} with a recurrent neural network (RNN; a sequential ANN) f_{RNN}. This allows the shape of the human learning rule to be learned empirically. Our next models expand the list of inputs to the value update f_{RNN} to assess which information humans use to calculate values $v(a)$. For example, are value updates contextualized by the current value landscape, or is the previous value of an action enough to determine its new value? Lastly, we test if other processes besides reward-based learning affect human choice. We construct an ANN that calculates

Table 1: Specification of all models in terms of their update and action rules. "Model Names" (left column) are structured as follows: "$RL_{\alpha\beta}$" refers to the basic, classic RL model. Model names containing "RNN" ["LSTM"] refer to models that replace the classic RL update rule with a flexible RNN [LSTM]: Their update rules (right column) use f_{RNN} [f_{LSTM}] rather than f_{RL}. "O-", "S-", and "OS-" (left column) indicate the memory-based inputs to each model's update rule (right column): "O-" and "OS-" models have access to their own previous output [v, c, or g]; "S-" and "OS-" models have access to their previous hidden state s. "-v" and "-c" (left column) indicate the observation-based inputs to each model's update rule (right column): rewards r allow for the calculation of values v; and choices a for choice-kernels c. "-cv" refers to models that track both v and c, and combine them at decision time: The action rule (right column, top row in each panel) in "v" ["c"] models is based on just values v [choice kernels c]; "cv" models calculate both and combine them additively (i.e., without the capability to show c-v interactions). "-all" indicates models that have access to all available information simultaneously, allowing them to model c-v interactions. Best-fitting models are in bold.

Model Names	**Update and Action rules**
Value ("v") Models	$p_t(a) = softmax(v_t(a))$
$RL_{\alpha\beta}$	$v_{t+1}(a_t) \mathrel{+}= f_{RL}(v_t(a_t), r_t)$
RNN-v	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t)$
O-RNN-v	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t, v_t)$
S-RNN-v	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t, s_t)$
OS-RNN-v	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t, v_t, s_t)$
LSTM-v	$v_{t+1}(a_t) \mathrel{+}= f_{LSTM}(v_t(a_t), r_t, s_t)$
Choice-Kernel ("c") Models	$p_t(a) = softmax(c_t(a))$
RNN-c	$c_{t+1}(a) = f_{RNN}(a_t)$
O-RNN-c	$c_{t+1}(a) = f_{RNN}(a_t, c_t)$
S-RNN-c	$c_{t+1}(a) = f_{RNN}(a_t, s_t)$
OS-RNN-c	$c_{t+1}(a) = f_{RNN}(a_t, c_t, s_t)$
LSTM-c	$c_{t+1}(a) = f_{LSTM}(a_t, c_t, s_t)$
Additive ("cv") Models	$p_t(a) = softmax(v_t(a) + c_t(a))$
RNN-cv	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t)$ $c_{t+1}(a) = f_{RNN}(a_t)$
O-RNN-cv	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t, v_t)$ $c_{t+1}(a) = f_{RNN}(a_t, c_t)$
S-RNN-cv	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t, s_t)$ $c_{t+1}(a) = f_{RNN}(a_t, s_t)$
OS-RNN-cv	$v_{t+1}(a_t) \mathrel{+}= f_{RNN}(v_t(a_t), r_t, v_t, s_t)$ $c_{t+1}(a) = f_{RNN}(a_t, c_t, s_t)$
LSTM-cv	$v_{t+1}(a_t) \mathrel{+}= f_{LSTM}(v_t(a_t), r_t, s_t)$ $c_{t+1}(a) = f_{LSTM}(a_t, c_t, s_t)$
Interactive ("all") Models	$p_t(a) = softmax(g_t(a))$
$RL_{\alpha\beta pf}$	$g_{t+1}(a_t) \mathrel{+}= f_{RL}(v_t(a_t), r_t, a_t)$
RNN-all	$g_{t+1} = f_{RNN}(r_t, a_t)$
O-RNN-all	$g_{t+1} = f_{RNN}(r_t, a_t, g_t)$
S-RNN-all	$g_{t+1} = f_{RNN}(r_t, a_t, s_t)$
OS-RNN-all	$g_{t+1} = f_{RNN}(r_t, a_t, g_t, s_t)$
LSTM-all	$g_{t+1} = f_{LSTM}(r_t, a_t, s_t)$

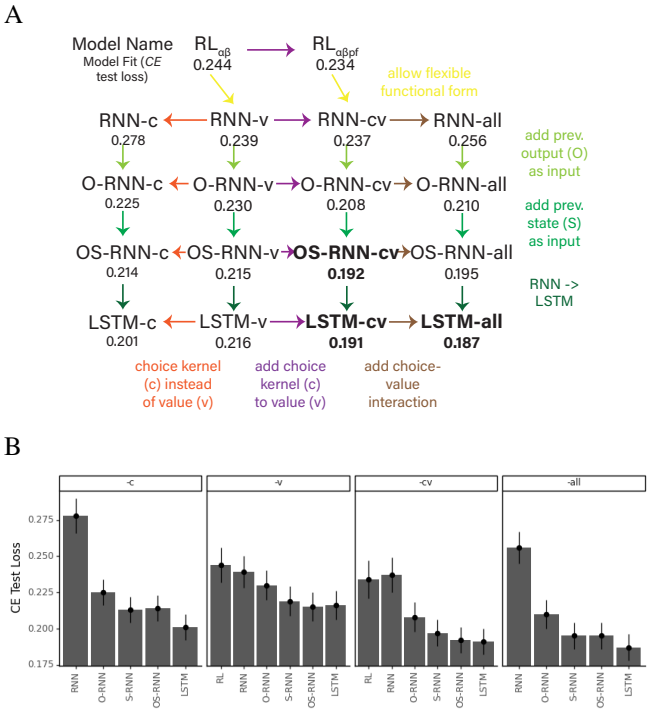

Figure 1: Overview and fit of the main models. A) Models are arranged according to their complexity, starting with $RL_{\alpha\beta}$ at the top-left, and ending with LSTM-all at the bottom-right. Three main features differentiate $RL_{\alpha\beta}$ from LSTM-all: The ability to flexibly learn the update rule from data (yellow, diagonal); additional memory capacity (green hues, vertical); and the existence of a reward-blind choice mechanism, and its interaction with the reward-based mechanism (red hues, horizontal). Arrow colors show which mechanism is added for each model. Numbers underneath model names indicate model fit (explained in panel B). Best-fitting models in bold. B) Model fits, i.e., *CE* losses on the held-out test dataset (lower is better). Each panel shows one model family ("-c": Choice-kernel models; "-v": value models; "-cv": additive models; "-all": interactive models; see Table 1).

reward-independent "choice kernels", and explore the interactions between value and choice kernel.

Classic Cognitive Models: "$RL_{\alpha\beta}$" and "$RL_{\alpha\beta pf}$" RL models are built on the notion of values and reward prediction errors (Daw, 2011; Wilson and Collins, 2019): Learners acquire, through trial-and-error, a value $v(a_i)$ for each action a_i. Values are acquired incrementally, by updating them each time a reward r is observed, based on the reward prediction error $rpe = r - v(a_i)$. To update $v(a_t)$, having observed r_t, previous values are moved slightly in the direction of the observed outcome: $v_{t+1}(a_t) = v_t(a_t) + \alpha * rpe$, with the learning rate parameter α controling the update size. To select actions based on these (potentially continuously evolving) values, the cognitive literature often employs the "softmax" transform: $p(a) = softmax(\beta * v(a))$.[1] The inverse decision temperature parameter β controls the stochasticity.

To fit our $RL_{\alpha\beta}$ model to the dataset, we initialized action values at 0.5 and performed gradient descent (Adam optimizer) on the cross-entropy loss $CE = -\sum_{s=1}^{s_{max}} \sum_{t=1}^{t_{max}} a_{1h} * log(p(a))/(p_{max} + t_{max})$. a_{1h} is the one-hot vector of participants' actions; $p(a)$ are model-derived action probabilities for each subject s and trial t (batch size $s_{max} = 64$). Parameters $\alpha = 0.77$ and $\beta = 7.05$ led to the best fit on the training data. (For testing loss and prediction accuracy on held-out participants, see Table 1 and Fig. 1B).

To assess qualitative model fit (Palminteri et al., 2017), we characterized models' ability to reproduce human behavior: We simulated behavior from the best (in training) among four models, asking the model to select actions based on its own proposed action probabilities, using human-fitted parameters. We simulated $n = 825$ agents with $t = 150$ trials, and subjected the simulated agents to the same behavioral analyses as humans. Although the $RL_{\alpha\beta}$ model learned to perform the task, it did not reproduce typical human behavioral patterns (Fig. 2B). To address this shortcoming, the $RL_{\alpha\beta}$ model is commonly expanded, for example by forgetting and choice persistence mechanisms, parameterized by f and p, respectively (e.g., Eckstein et al., 2022). Adding these parameters in model $RL_{\alpha\beta pf}$ led to a slight improvement in model fit (Fig. 1).

Most Flexible Model: "LSTM-all" We next fit the most general ANN: a long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) model we call "LSTM-all" because it has access to "all" observable trial information r_t and a_t. We trained LSTM-all to output a 4-dimensional "gist" vector $g_t(a)$ on each trial t, which—similar to $RL_{\alpha\beta}$'s $v_t(a)$—was submitted to the softmax function to obtain action probabilities (Table 1).

LSTM-all was fitted using the same CE loss as $RL_{\alpha\beta}$ and $RL_{\alpha\beta pf}$. Hyperparameters were chosen based on visual inspection of training trajectories, and then fixed for all mod-

[1] We use the notation a to signify the vector of all actions; a_t refers to the action chosen at trial t. $p(a)$ [$v(a)$] denotes the vector of all action probabilities [values].

els.[2] LSTM-all's behavioral validation was outstanding, with all human behavioral markers reproduced reliably (Fig. 2B; for quantitative fit, see Fig. 1B). Hence, as expected, the constrained $RL_{\alpha\beta}$ model does not capture human behavior as well as the flexible LSTM-all. This raises the question which specific mechanisms LSTM-all capture that $RL_{\alpha\beta}$ was lacking.

Basic Value RNN: "RNN-v" We first tested whether relaxing the functional form of the $RL_{\alpha\beta}$ value update improves model fit. We created RNN-v, a model identical to $RL_{\alpha\beta}$ except that it replaced f_{RL} with an unconstrained RNN f_{RNN}. Notably, f_{RNN} was constrained to the same inputs as f_{RL} (i.e., did *not* recurrent states like regular RNNs; Table 1). Specifically, RNN-v implements a recurrent 3-layer MLP with 2 input units $(v_t(a_t), r_t)$, 16 hidden units (s_t), and 1 output unit $(u_t(a_t))$. RNN-v's free parameters θ_{RNN} contain weight matrices W and biases b. *tanh* is a non-linear activation function:

$$s_t = tanh(W_h[v_t(a_t), r_t] + b_h)$$
$$u_t(a_t) = W_u s_t + b_u$$
$$v_{t+1}(a_t) = v_t(a_t) + u_t(a_t)$$

As expected, RNN-v achieved better quantitative (Fig. 1B) and qualitative fit (Fig. 2B), though the differences were relatively small.

After training, the fitted weights θ_{RNN} of f_{RNN} implement a value update function entirely learned from human behavior (Fig. 2A, second row). Because f_{RNN} has the same inputs and outputs as f_{RL} (Table 1), the two are directly comparable (Fig. 2A, top row). We found that when the chosen action a_t had a high value prior to the update (x-axis), f_{RNN} and f_{RL} were similar: Larger rewards (yellow) resulted in larger updated values than smaller rewards (blue). There was a difference with respect to reward sensitivity, however: Whereas f_{RL} showed an even spread over reward magnitudes, f_{RNN} was compressed at the high and low ends. Furthermore, when the chosen action a_t had a low value prior to the update, f_{RNN} diverged substantially from f_{RL}: Previously low values stayed low, irrespective of the reward magnitude. In sum, allowing functional flexibility in the value update improved RNN-v's fit to human data. The learned value update function showed a distortion in human reward sensitivity and reduced updating for low-valued actions. [3] However, compared to other mechanisms explored later, the improvement in RNN-v's fit was relatively small (Fig. 1B).

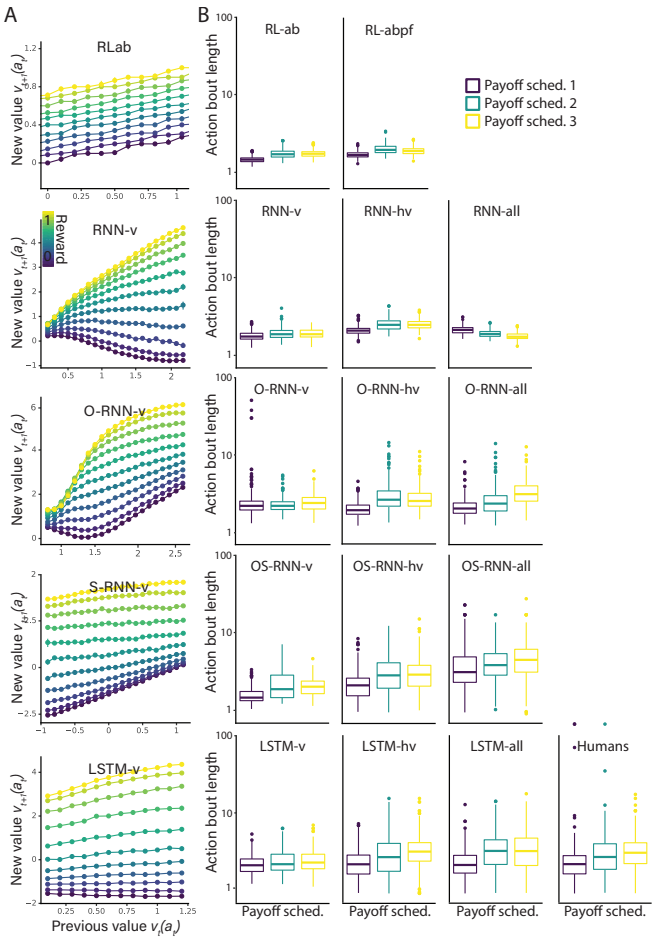

Figure 2: A) Analyzing learned learning rules. All "v models" update $v(a)$ on each trial, using parameters and (except for $RL_{\alpha\beta}$) update functions (learning rules) learned from human data. X-axes $(v_t(a_t))$ and colors (r_t) show function inputs, y-axes show function outputs $(v_{t+1}(a_t))$, for a full specification of the learned functions. Methods: 10.000 random input tuples were passed through each update function to obtain outputs. r_t was sampled randomly across the allowed range $[0,1]$; $v_t(a_t)$ and s_t were sampled within 3 standard deviations of their mean along their first principal component (PC). Points indicate means, error bars standard errors of the mean, after binning data points into 0.1-sized bins based on reward magnitude. Differences in y-axis scale can be interpreted as differences in "decision temperature". B) Assessing qualitative model fit using behavioral simulations. We used human-fitted parameters θ to simulate artificial behavior, and subjected all (artificial and human) datasets to the same behavioral analyses. Action bout length (participants' average numbers of subsequent identical actions) is one example behavior. Box plots show the 25th, 50th (median), and 75th percentile over participants; whiskers extend to the 1.5-fold interquartile range; remaining participants are represented by dots. Models are arranged in the same way as in Fig. 1. Qualitative model fit, i.e., similarity to human behavior (top-right), closely mirrors quantitative model fit (Fig. 1, Table 1), with similar results for behaviors other than action bout length.

[2] batch size s_{max}: 64; number of training steps: 200.000; size of hidden layer: 16; Adam optimizer learning rate for $RL_{\alpha\beta}$, $RL_{\alpha\beta pf}$, and all models of the "-all" family: 0.001; for all other models: 0.0001

[3] Note that these patterns could conflate task-based and individual differences. For example, it is possible that some participants, e.g., due to a lack of task engagement, showed a lack of reward-informed action choice (i.e., no detectable value update in v_{t+1}) and hence observed generally smaller rewards (i.e., low values v_t). These participants might explain the left (low-value) side of the update function. A different group of participants, showing optimal task engagement, might however show great sensitivity to reward differences (i.e., easily detectable value update in v_{t+1}), and hence observe consistently large rewards (i.e., high values v_t). These participants might explain the right (high-value) side of the update function.

Value RNN with Output Memory: "O-RNN-v" We next investigated whether the ability to access not just the value of the chosen action $v_t(a_t)$, but of all actions $v_t(a)$ improves model fit, creating model O-RNN-v (Table 1). Access to $v(a)$ allows O-RNN-v to condition individual value updates on its own, as well as all other actions' values. Indeed, model fit improved (Fig. 1B, Fig. 2B): In analyses not shown here, we observed that O-RNN-v learned to apply different value updates in different value contexts. This suggests that in the current task, humans conditioned individual value updates on the current value landscape, rather than performing them in isolation, in accordance with previous findings (e.g., Palminteri et al., 2015).

Value RNN with State Memory: "OS-RNN-v" We next investigated the effects of memory more broadly, constructing OS-RNN-v, which receives its own prior hidden state s_t as an additional input, completing the classic RNN architecture (Table 1). Access to s_t allows the model to carry information between trials that is not directly related to action choice (in the way $v(s)$ is), and can hence capture long-term contingencies or action plans (or individual differences, see Discussion). This change further improved quantitative (Fig. 1) and qualitative fit (Fig. 2B).

Value LSTM: "LSTM-v" We then constructed LSTM-v, a vanilla LSTM trained on the value-based inputs (Table 1): Like the other models in the "v family", LSTM-v submitted a 4-dimensional vector $v(a)$ to the choice rule, and provided updates $u_t(a_t)$ to the previous action (a_t). LSTM-v is the most powerful model in the v-family, with a potentially improved capacity to capture long-term dependencies. However, model fit did not improve (Fig. 1, Fig. 2B).

Choice-Kernel RNNs: "-c Models" Despite access to the full capacity of the LSTM architecture, LSTM-v does not achieve the same fit as LSTM-all (Fig. 1, Fig. 2B). The reason is that LSTM-v lacks access to the identity of the chosen action a_t (Table 1). We therefore investigated next how much behavioral variance in the human data could be explained by models based on action identity a_t. To this aim, the choice-kernel or "c family" of models had access to a_t, but *not r* (Table 1). We constructed the c family in direct correspondence to the v family, allowing direct comparison of the role of memory (differences between "O-", "S-", and "OS-" models).

We found that similar to value RNNs, choice-kernel RNNs benefit from access to c ("O" models) and s ("S" models"), with even larger observed differences in model fit (Fig. 1). This suggests that long-term memory might play a larger role in action-history dependent (i.e., choice kernel-based) than in reward-dependent (i.e., value-based) choice.

Interestingly, the choice kernel-based model with full memory capacity (LSTM-c) explains more variance than its value-based partner LSTM-v (Fig. 1), suggesting that in the current task, participants relied more on action patterns than reward information. This could reflect an increased use of exploratory strategies (Fintz et al., 2021) or lack of task engagement.

Additive Combination of Value and Choice Kernel: "■-cv Models" We next asked how v and c processes interact. We first constructed models that calculate both choice kernels and values independently, and only (additively) combine them at the final decision stage ("additive" family, Table 1). Indeed, additive models fit the human data much better than either component model alone, at every level of memory use (Fig. 1). Additive models also show substantially better qualitative fit than prior models (Fig. 2B).

Interactions between Value and Choice Kernel: "-all Models" The models in the additive family lack one crucial capability of LSTM-all, notably the non-linear combination of reward and action identity: By design, additive models cannot capture processes that combine, compare, or contextualize reward and action identity with each other because the calculations are performed separately (Table 1).

We hence constructed the final "interactive" model family to assess evidence for choice kernel-value interactions (Table 1). Interestingly, interactive models do not generally provide improvements in quantitative (Fig. 1) or qualitative fit (Fig. 2B) over additive models (with the potential exception of LSTM-all). This might suggest that humans process reward histories and action histories using distinct cognitive—and potentially neural—pathways, in accordance with previous findings (K. J. Miller et al., 2019).[4]

Discussion

This paper addresses several open questions in the field of RL and cognitive modeling: 1) How much behaviour in typical RL tasks is explained by theoretical, hand-crafted RL models? Echoing previous findings in RL (e.g., Dezfouli, Ashtiani, et al., 2019; Dezfouli, Griffiths, et al., 2019; Fintz et al., 2021; Jaffe et al., 2022; Song et al., 2017; Yang et al., 2019) and other fields (e.g., Agrawal et al., 2020; Battleday et al., 2020; Kuperwajs et al., 2022; Peterson et al., 2021), we find that hand-crafted, cognitive RL models leave substantial amounts of behavioral variance unexplained, which can reliably be captured using more flexible, ANN-based models.

2) If hand-crafted RL does not explain human choice, which mechanisms do? We find strong evidence for an outcome-insensitive decision mechanism, which might capture exploration or habits (K. J. Miller et al., 2019). This mechanism can be inspected and interpreted based on input-output mappings just like the value-based mechanism (Fig. 2A), though we are not showing the results here. Surprisingly, we find no evidence for non-linear interactions between reward-based and outcome-insensitive mechanisms, suggesting separate cognitive mechanisms.

[4]Note that this conclusion is based on the assumption that the current dataset was large enough to fit all models optimally, which might not be the case (see Discussion and Fig. 3). Some of the observed pattern could also arise if more complex models were underfit.

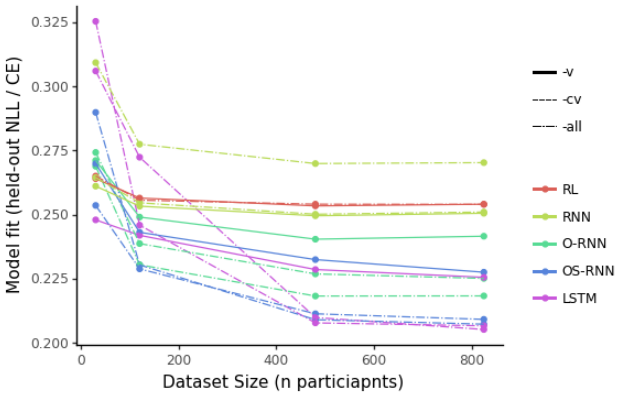

Figure 3: Model fits vary with training data size. Each line refers to one model, colors and linetypes specify model categories. Each model was fitted to four different datasets (30, 120, 480, or 825 participants) for 200.000 training steps, using the same hyperparameters. The trained models were then tested on the same held-out data (y-axis, negative log-likelihood / cross-entropy). More flexible models (e.g., LSTM, OS-RNN) showed greater improvements in fit than less flexible models (e.g., $RL_{\alpha\beta}$), driven by both increased overfitting (i.e., worse fit to held-out data) when trained on smaller datasets, and superior generalization (i.e., better fit to held-out data) when trained on larger datasets.

3) How do humans perform RL? RL theory specifies a particular updating rule, but other functional forms are possible. Our framework shows which form the human update takes when learned directly from data.

4) What role does memory play in human action selection? While the literature on memory and its interactions with RL is rich, it has traditionally been difficult to create hand-crafted models that encompass both processes (Collins, 2019; Daw and Shohamy, 2008). We show distinct roles of memory for reward-based and reward-insensitive choice processes.

These findings extend previous studies that have aimed to achieve interpretability when using ANNs as cognitive models. Previous suggestions range from shifting the research focus from explanation to prediction (Yarkoni and Westfall, 2017), analyzing ANNs' hidden representations (Kriegeskorte, 2015; Schaeffer et al., 2020) or behavioral predictions (Agrawal et al., 2020; Dezfouli, Griffiths, et al., 2019; Fintz et al., 2021), directly comparing ANNs to hand-crafted cognitive models (Fintz et al., 2021), creating ANNs that expose interpretable information in their hidden state (Dezfouli, Ashtiani, et al., 2019; Dezfouli, Griffiths, et al., 2019), or integrating ANN-based representations (Battleday et al., 2020; Noti et al., 2016) or entire computational components into existing, hand-crafted models (Peterson et al., 2021). Our approach is very closely related to Peterson et al. (2021).

Limitations and Future Work

Dataset Size One limitation of the current study is the dataset size. Even though at 965 participants, the dataset (Bahrami and Navajas, 2020) is an order of magnitude larger than most standard studies in the field, we observed indices that it might be too small: We calculated the test loss of our models after training on different data sizes. Our most flexible models still showed improvements leading up to the largest size (Fig. 3). Furthermore, some results were sensitive to small changes in the size of training and testing data or on participants' assignments. We see two main solutions: collecting a larger dataset or reducing model complexity in accordance with the given dataset (e.g., by reducing the number of hidden neurons).

Task Design Some of our results might be specific to the current task, rather than pertain to learning and decision making more broadly. For example, the strong role of choice kernel-based control might be related to the relatively large number of four alternatives, or the common closeness of alternatives to each other, which makes it difficult to identify the best (Fintz et al., 2021). If this is the case, conclusions drawn from this dataset might not explain the fundamental set-up of human learning, but rather elucidate particular task strategies (Eckstein et al., 2021; Nussenbaum and Hartley, 2019).

To alleviate this problem, future work needs to exhaustively sample the space of task features over which conclusions are aimed to be drawn (e.g., number of bandits; stochastic or deterministic reward scheme; volatile or stable trial structure; correlated, anticorrelated, uncorrelated bandits; etc.). Such a dataset would allow to marginalize over individual task features (Peterson et al., 2021), and follow a ubiquitous aspiration to introduce richer, more complex paradigms into the study of cognition (Battleday et al., 2020; Ma and Peters, 2020).

Individual Differences Like the majority of cognitive ANN studies, our approach does not explicitly model individual differences (however, see Dezfouli, Ashtiani, et al., 2019; Dezfouli, Griffiths, et al., 2019). This is particularly relevant in the current setting because our basic RL models have no capacity to capture individual differences, while more advanced models have this capacity as long as they are endowed with state-memory: These advanced models can adapt to individuals by conditioning decisions on their hidden state, which can represent unique individual characteristics.

Conclusion

In this paper, we introduce "hybrid ANNs", a combination of classic RL models and ANNs. These models have both predictive (excellent fit) and explanatory power (interpretability of the resulting models), and potentially provide detailed insight into human cognitive processes.

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via Scientific Regret Minimization

[Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *117*(16), 8825–8835.

Bahrami, B., & Navajas, J. (2020). 4 Arm Bandit Task Dataset [Publisher: OSF].

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 5418.

Box, G. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics* (pp. 201–236). Elsevier.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling*. Sage Publications, Incorporated.

Collins, A. G. E. (2019). Reinforcement learning: Bringing together computation and cognition. *Current Opinion in Behavioral Sciences*, *29*, 63–68.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII*.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans [Number: 7095 Publisher: Nature Publishing Group]. *Nature*, *441*(7095), 876–879.

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593–620.

Dezfouli, A., Ashtiani, H., Ghattas, O., Nock, R., Dayan, P., & Ong, C. S. (2019). Disentangled behavioural representations. *Advances in Neural Information Processing Systems*, *32*.

Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P., & Balleine, B. W. (2019). Models that learn how humans learn: The case of decision-making and its disorders. *PLOS Computational Biology*, *15*(6), e1006903.

Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2022). Reinforcement learning and bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience*, 101106.

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137.

Fintz, M., Osadchy, M., & Hertz, U. (2021). *Using Deep Learning to Predict Human Decisions, and Cognitive Models to Explain Deep Learning Models* (preprint). Neuroscience.

Ger, Y., Shahar, M., & Shahar, N. (2023). Using recurrent neural network to estimate irreducible stochasticity in human choice-behavior.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Jaffe, P. I., Poldrack, R. A., Schafer, R. J., & Bissett, P. G. (2022). Discovering dynamical models of human behavior.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446.

Kuperwajs, I., Schuett, H., & Ma, W. J. (2022). Improving a model of human planning via large-scale data and deep neural networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning [Number: 7553 Publisher: Nature Publishing Group]. *Nature*, *521*(7553), 436–444.

Ma, W. J. (2019). Bayesian Decision Models: A Primer. *Neuron*, *104*(1), 164–175.

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior* (tech. rep. arXiv:2005.02181) [arXiv:2005.02181 [cs, q-bio] type: article]. arXiv.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without Values. *Psychological review*, *126*(2), 292–311.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*(3), 363–386.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, *11*, 49–54.

Navarro, D. J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior*, *2*(1), 28–34.

Noti, G., Levi, E., Kolumbus, Y., & Daniely, A. (2016). Behavior-Based Machine-Learning: A Hybrid Approach for Predicting Human Decision Making [arXiv:1611.10228 [cs]].

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, *40*, 100733.

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *6*(1), 8096.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation*, *20*(4), 873–922.

Schaeffer, R., Khona, M., Meshulam, L., Laboratory, I. B., & Fiete, I. R. (2020). Reverse-engineering Recurrent Neural Network solutions to a hierarchical inference task for mice [Pages: 2020.06.09.142745 Section: New Results].

Song, H. F., Yang, G. R., & Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 24.

Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *11*(1), 3574.

Sutskever, I., & Nair, V. (2008). Mimicking Go Experts with Convolutional Neural Networks. In V. Kůrková, R. Neruda, & J. Koutník (Eds.), *Artificial Neural Networks - ICANN 2008* (pp. 101–110). Springer.

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data (T. E. Behrens, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *8*, e49547.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.