Understanding Atypical Behavior in Human Decision Making via Recurrent Neural Network Interpretability

Jin Zida *
Department of Computer Science
University of Science and Technology of China
Hefei, Anhui, China
rie.acad@gmail.com

Li Ji-An *
Neurosciences Graduate Program
University of California San Diego
La Jolla, CA 92093
jil095@ucsd.edu

Marcelo G. Mattar
Department of Psychology
New York University
New York, NY 10012
marcelo.mattar@nyu.edu

Abstract

As artificial intelligence continues to revolutionize scientific research, its application in understanding decision-making behavior in neuroscience and cognitive science stands out. Specifically, artificial neural networks have shown the potential to produce more accurate predictions than classical cognitive models in tasks related to the behavior of biological agents. However, uncovering the underlying mechanisms of these networks is challenging due to their limited interpretability. Here, we studied the computational mechanisms of recurrent neural networks (RNNs) trained on human behavior in a two-armed bandit task featuring rare rewards. Interpreting these RNNs within a dynamical systems framework, we discovered numerous strategies underlying heterogeneous, atypical behavioral patterns that evade classical cognitive modeling, such as a tendency to shift actions after receiving a reward. By utilizing features derived from tiny RNNs that encapsulate these strategies, we achieved a diagnostic accuracy comparable to that obtained from the full, black-box RNNs. Overall, by introducing innovative analysis methods based on RNN interpretability, our work establishes a connection between explainable artificial intelligence and computational psychiatry.

1 Introduction

Understanding adaptive behavior and decision-making stands as a cornerstone in the fields of neuroscience and cognitive science. To characterize the underlying computational mechanisms of adaptive behavior, researchers often develop cognitive models based on normative principles such as reinforcement learning [1, 2]. Artificial neural networks have recently emerged as potent modeling tools, often surpassing classical cognitive models in their ability to offer more precise predictions of the behavior of biological agents [3–5]. However, the impressive predictive performance of these networks is shadowed by a major challenge – their inherent limited interpretability constrains the depth of insights into underlying mechanisms we can extract.

In this study, we examined the emergent mechanisms of recurrent neural networks (RNNs) trained on behavioral data from human subjects performing a two-arm bandit task featuring rare rewards

*Equal contribution

Under review.

[3]. We identified limitations in the prior study on similar RNNs trained with the same behavioral data [3]: (1) the RNNs, fitted on group data, were treated as one virtual subject, while the individual differences within each diagnostic group were overlooked; (2) the RNNs were primarily studied through simulated behavior, i.e., using predefined input sequences of actions and rewards to analyze the output probabilities. However, these input sequences may not represent the extensive input space, and circuit-level mechanisms are yet to be identified.

Addressing these issues, we analyzed the underlying mechanisms of these RNNs for each individual subject using a recent interpretative framework based on dynamical systems [4]. We discovered heterogeneous cognitive strategies among subjects, indicating diverse individual differences. These strategies thus provided explanations for atypical behavioral patterns observed in subjects, such as a tendency to shift actions after receiving a reward. We further extracted features of these strategies using tiny RNNs and found that these features can yield a diagnostic prediction performance on par with the larger, black-box RNNs from the prior study [3].

2 Task and model fitting

101 subjects completed a reward learning task (Fig. 1a): 34 healthy (control) subjects, 34 depression subjects, and 33 bipolar subjects. In each trial, subjects chose between a left action (A_1) or a right action (A_2) to earn rewards. Subjects entered their choices at their own pace. In each block, the better action had a reward probability of 0.08, 0.125, or 0.25, whereas the worse action had a 0.05 probability. Each subject completed 12 blocks, with an average of 109 trials per block.

We fitted RNNs to data from all subjects (Fig. 1b), using five-sixths of the data as a training set, and evaluating the fitted models on the remaining one-sixth (i.e., a six-fold cross-validation). The input layer of the RNN consists of the action and the reward from the current trial. The recurrent layer employs the gated recurrent unit (GRU) with 20 neurons. The output action probability is derived by sending the recurrent layer activations through a fully connected layer, then a softmax layer, and compared to next-trial choices using cross-entropy loss.

We found that the performance (a loss of 0.271) of our GRU-based networks matched the predictive performance of previously studied LSTM models [3]. They also significantly outperformed classical cognitive models like Q-learning (QL; also see Table 2 in the following sections). By augmenting the RNNs with an input corresponding to the subject embedding [6], we found that the loss reached 0.258, further improved by 0.013, suggesting that knowledge of the subject identity improves the predictions of the network. Accordingly, the results of Group-GRU models reported in all subsequent sections correspond to those augmented with the subject embedding layer.

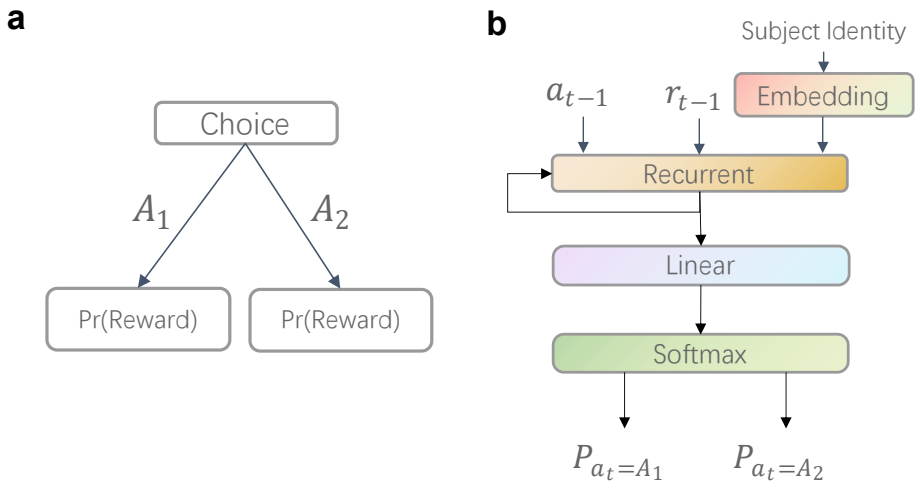

Figure 1: (a) The task structure. The subjects choose between two actions. (b) The recurrent neural network with the subject embedding layer.

3 Understand RNNs via the logit analysis

Given the superior accuracy of these RNNs over classical cognitive models, we next aimed to identify the emergent mechanisms acquired by these RNNs, i.e., cognitive strategies employed by these subjects. We adopted the dynamical systems approach developed by Ji-An et al. [4] to understand the solutions learned by these RNNs. This approach analyzes the temporal changes in state variables (i.e., neuronal activations) as a function of inputs and state variables. A notable special case is the logit analysis, which projects state variables and their temporal changes onto the network's output space.

Formally, for trial t, the logit $L(t)$ produced by a model is defined as $L(t) = \log[\Pr_t(A_1)/\Pr_t(A_2)]$, where $\Pr_t(A)$ represents the probability of selecting action A. A positive logit indicates a preference for A_1 over A_2 and vice versa. The logit-change $\delta L(t)$ is defined as $L(t+1) - L(t)$, the difference between logits of two consecutive trials. A positive (negative) logit-change indicates that the agent's preference for A_1 (A_2) increases (decreases) due to the input at trial t (i.e., the reward following the action).

For each subject, we plotted the logit-change against logit for each trial provided by the group-GRU, categorizing by the four input conditions ($[A_1\ R = 0]$, $[A_1\ R = 1]$, $[A_2\ R = 0]$, $[A_1\ R = 1]$). Using this approach, we observed a remarkable degree of individual variability in logit patterns, with different subjects manifesting different strategies. Below we outline representative strategies manifested by example subjects:

(i) Switching strategy (Fig. 2a, top): The logit-change δL lies above the diagonal line ($\delta L = -L$) for light/dark red points (A_2) with $L < 0$ and below the diagonal line for light/dark blue points (A_1) with $L > 0$, suggesting a logit sign reversal in the subsequent trial (e.g., if $L(t) < 0$ and $\delta L(t) > -L(t) > 0$, then $L(t+1) = L(t) + \delta L(t) > 0$). Therefore, the subject typically switches the action preference each trial, a pattern confirmed in the raw data. The acquisition of this strategy by the networks explains the oscillatory behavior previously observed in such RNNs [3], i.e., initial oscillations (observing action sequence "..., A_1, A_2, A_1, A_2") can trigger the following action switches (predicting action sequence "A_1, A_2, A_1, ...").

(ii) Inverted-reward strategy (Fig. 2a, middle): In classical model-free RL models (e.g., Q-learning model), a reward following action A_1 always leads to a positive logit change (thus, the agent prefers A_1 more in the next time), while no reward following A_1 leads to a negative logit change (e.g., see [4]). For behavior generated by such model-free agents (see this Q-learning agent fitting to the same subject in Fig. 2c, middle), we should expect that the trials of $[A_1\ R = 1]$ (dark blue) lie above trials of $[A_1\ R = 0]$ (light blue), and similarly, trials of $[A_2\ R = 1]$ (dark red) lie below trials of $[A_2 R = 0]$ (light red). Strikingly, we found that this is not the case in most subjects. In contrast, a reward following action A_1 usually leads to a negative logit change (dark blue points below the x-axis), decreasing the preference for action A_1, suggesting a tendency to switch. The lack of reward following action A_1 (light blue points around/above the x-axis) usually leads to a zero or positive logit change, indicating a preference to stay on the same action. This inverted role of reward, which has not been reported in the literature, explains the puzzling phenomenon that a reward causes a dip in the RNN's output probability of staying on the same action [3].

(iii) Temporal mixture of strategies (i) and (ii) (Fig. 2a, bottom), e.g., using the switching strategy for several trials (light red points for A_2 with $L < 0$ above the diagonal line and light blue points for A_1 with $L > 0$ below the diagonal line) and then alternating to the inverted-reward strategy (light red points around/below the x-axis and light blue points around/above the x-axis). These alternations of strategies were consistent with the inspection of raw choice data. Similar phenomena of alternating strategies in perceptual decision-making in rodents have been reported [7].

In sum, we found a variety of cognitive strategies exhibited by different subjects. This implies that networks trained on group data don't emulate a singular group representative. Instead, they simulate diverse strategies, tailoring their predictions to each subject's identity and distinct history.

4 Diagnostic label prediction

Next, we aimed to determine if a subject's logit patterns contain information about their healthy and pathological status, a core objective for computational psychiatry. Specifically, we investigated whether these strategies can predict diagnostic labels. As an initial step, we trained one-dimensional

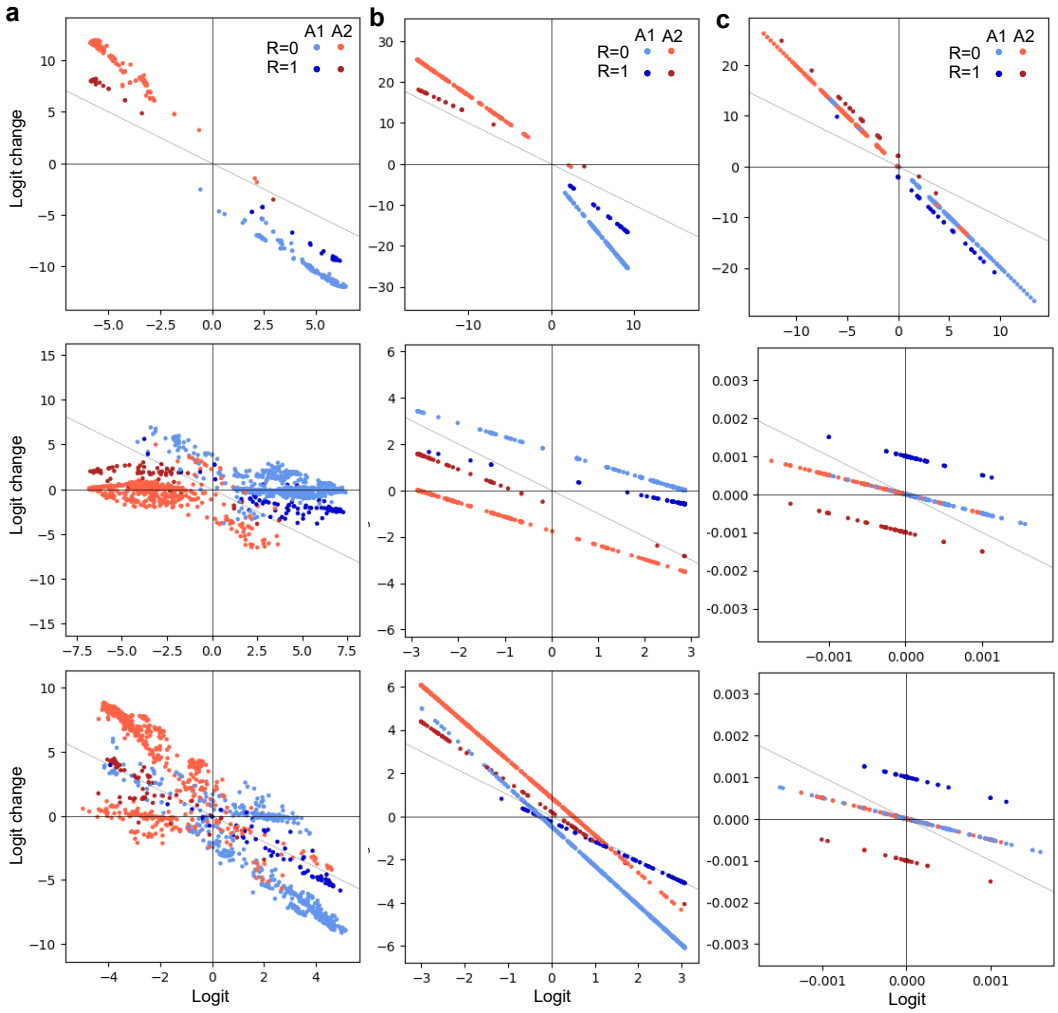

Figure 2: The logit patterns produced by (a) Group-GRU, (b) Ind-SLRU with one unit, and (c) the one-dimensional Q-learning model (classical cognitive model). (Top) Switching strategy (i) from one example subject. (Middle) Inverted-reward strategy (ii) from one example subject. (Bottom) Temporal mixture of strategies (iii) from one example subject. The diagonal line is defined as $\delta L = -L$.

RNNs (technically, switching linear recurrent unit, or SLRU), which acted as a first-order approximation of a one-dimensional compression of these strategies.

In an SLRU, the hidden state h_t ($t > 0$) is updated as follows [4]:

$$h_t = W^{(x_{t-1})} h_{t-1} + b^{(x_{t-1})} \tag{1}$$

where $W^{(x_{t-1})}$ and $b^{(x_{t-1})}$ are the weight matrices and biases selected by the input x_{t-1}. When h_t is one-dimensional, each $W^{(\cdot)}$ and $b^{(\cdot)}$ are scalars. Importantly, these weights are directly related to the learning rate ($\alpha = 1 - W$) in the RL models, while the biases are similar to the reward term ($= \alpha r$).

Fitting these one-dimensional SLRUs to individual choice data, we found that they achieve an average loss of 0.337. While this is worse than the Group-GRU model (0.258), but significantly outperforms the one-dimensional QL model (i.e., $V(A_1, t+1) = (1-\alpha)V(A_1, t) + \alpha r \cdot \text{Sgn}(a_t = A_1)$) fitted to individual choice data (Table 2), suggesting they provide a reasonably well one-dimensional characterization of behavior. As expected, these one-dimensional SLRUs are effective at summarizing the dynamics of (i) and (ii) strategies (comparing Fig. 2b and a, top and middle) and less effective at dealing with the dynamics of the temporal mixture (iii) strategy (Fig. 2b, bottom). As a comparison, these one-dimensional QL models cannot effectively summarize these underlying

Dimensionality	Model	Cross-entropy loss
1	Ind-QL	0.555
1	Ind-SLRU	0.337
1	Ind-GRU	0.347
2	Ind-GRU	0.303
3	Ind-GRU	0.294

Table 1: Models fitted on the choice data from each subject (evaluated with cross-validation; reported results averaged over subjects). "Ind" represents the individual level.

Features & Classifier	Accuracy %
1-D Ind-QL + SVM	22
1-D Ind-SLRU + SVM	52
LSTM trained on each group[3]	52

Table 2: Diagnostic prediction performance (evaluated with leave-one-subject-out cross-validation).

strategies (comparing Fig. 2c and a). We thus extract these interpretable weights and biases from these SLRUs (one weight and one bias for each trial condition, effectively adjusted by the readout layer weights) to serve as diagnostic prediction features.

We employed nested cross-validation to evaluate the prediction of the diagnostic labels from these features, same as the procedure in the prior study [3]. In the outer loop, one subject was left out as the test dataset, and classifiers were trained on the remaining data with hyperparameters tuned using cross-validation. We found that the support vector machine (SVM) with a linear kernel achieves an overall 52% accuracy (for the confusion matrix, see Fig. 3), the best among all classifiers tested (random forest, AdaBoost, K nearest neighbor, logistic regression, and SVM). This is substantially better than the overall 22% accuracy of classification using the features from the QL model (i.e., learning rate and inverse temperature) and the 34% chance-level accuracy. Our approach differs from the original study that also achieved an overall 52% accuracy [3], where authors trained separate LSTMs for each group (healthy, depression, and bipolar) and assigned the label to the left-out subjects based on which LSTM produces the lowest loss. Our results indicate that the classification using interpretable features extracted from one-dimensional SLRUs is comparable to those using larger, black-box RNNs. This also suggests that interpretable features extracted from higher-dimensional SLRUs might even surpass the current prediction performance.

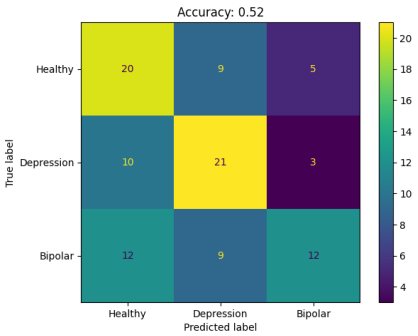

Figure 3: The confusion matrix of classification based on the features extracted from the one-dimensional SLRU.

5 Conclusion and future work

In this study, we harnessed the capabilities of recurrent neural networks to delve deeper into the intricacies of decision-making behavior. Through our interpretative approach, we can capture and

elucidate diverse cognitive strategies employed by subjects, most of which were previously overlooked by classical cognitive modeling. The use of tiny RNNs (SLRU) not only provided a more interpretable compression of these strategies, but also demonstrated impressive diagnostic prediction capabilities.

Looking ahead, several promising directions emerge for further exploration. First, we plan to train and analyze 2- or 3-unit SLRUs for each individual subject to capture richer features of cognitive strategies. Recognizing that the increased flexibility may pose challenges given the limited choice data from each subject, we will further leverage knowledge distillation techniques to transfer knowledge from larger group-trained GRU models to these smaller SLRUs [4]. We expect that features from these might enable better diagnostic prediction.

Second, we aim to detect and characterize the alternation of strategies more automatically and systematically by clustering RNN states across different trials. This allows us to determine whether the tiny RNNs (SLRUs) have the capacity to capture the alternation of strategies or whether other architectures (such as a mixture of SLRUs, similar to the mixture-of-experts architecture) are required.

These advancements would not only deepen our insights into decision-making behavior but could also pave the way for more effective explanatory methods in artificial intelligence and diagnostic tools in computational psychiatry.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] Anne Gabrielle Eva Collins. Reinforcement learning: Bringing together computation and cognition. *Current Opinion in Behavioral Sciences*, 29:63–68, 2019.

[3] Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, and Bernard W Balleine. Models that learn how humans learn: The case of decision-making and its disorders. *PLoS computational biology*, 15(6):e1006903, 2019.

[4] Li Ji-An, Marcus K Benna, and Marcelo G Mattar. Automatic discovery of cognitive strategies with tiny recurrent neural networks. *bioRxiv*, pages 2023–04, 2023.

[5] Kevin J Miller, Maria Eckstein, Matthew M Botvinick, and Zeb Kurth-Nelson. Cognitive model discovery via disentangled rnns. *bioRxiv*, pages 2023–06, 2023.

[6] Mingyu Song, Yael Niv, and Mingbo Cai. Using recurrent neural networks to understand human reward learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

[7] Zoe C Ashwood, Nicholas A Roy, Iris R Stone, International Brain Laboratory, Anne E Urai, Anne K Churchland, Alexandre Pouget, and Jonathan W Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, 2022.