Understanding atypical decision making behavior with recurrent neural networks

A major goal of cognitive science is to characterize the cognitive processes underlying healthy and pathological decision making. A traditional approach involves developing cognitive models grounded in normative principles such as reinforcement learning and Bayesian inference. These models are typically composed of few interpretable parameters that are fit to a subject's decisions and then used to characterize the differences between individuals or between groups. Recently, artificial neural networks have emerged as an alternative modeling framework, enabling better predictions of behavior and requiring less domain-specific knowledge than classical cognitive models. However, neural networks are notoriously difficult to interpret, and therefore seen as inadequate for characterizing distinct patterns of decision making. Here, we leverage tiny recurrent neural networks (RNNs) and dynamical-systems visualization shown to facilitate the identification of interpretable cognitive strategies in reward-learning tasks (Ji-An et al. 2023). We used this framework to analyze human decisions in a two-armed bandit task with rare rewards, contrasting healthy, depression, and bipolar subjects. Our research uncovered various strategies that drive diverse, atypical behavioral patterns missed by classical cognitive modeling, such as a tendency to shift actions after a reward. Utilizing features from tiny RNNs, we designed a diagnostic procedure that predicts an individual pathological status based solely on their decisions, with accuracy comparable to those from larger, black-box RNNs. Overall, our findings demonstrate the significant role of tiny RNNs and dynamical-systems interpretability in understanding individual differences in computational psychiatry.

We re-analyzed the data from Dezfouli et al. 2019, where 101 subjects (34 healthy, 34 depression, and 33 bipolar subjects) completed 12 blocks of a reward learning task with rare rewards (Fig. 1a). On each trial, subjects chose between a left (A_1) or a right action (A_2) at their own pace to earn rewards. In each block (109 trials on average), the better action had a reward probability of 0.08, 0.125, or 0.25, whereas the other had a probability of 0.05. We fitted an RNN (gated recurrent unit, GRU; 20 neurons) to predict next-trial choices from all subjects (cross-entropy loss), using the current-trial action and reward as the input (Fig. 1b). This GRU obtained a cross-validated test loss of 0.271, a performance comparable to that of LSTM models (Dezfouli et al. 2019) and substantially higher than classical cognitive models like Q-learning (QL; see below). This loss was further reduced to 0.258 (Fig. 1c) by augmenting the RNN with a subject embedding layer (model "Group-GRU"), as proposed in Song et al. (2021), suggesting that knowledge of subject identity improves the network predictions.

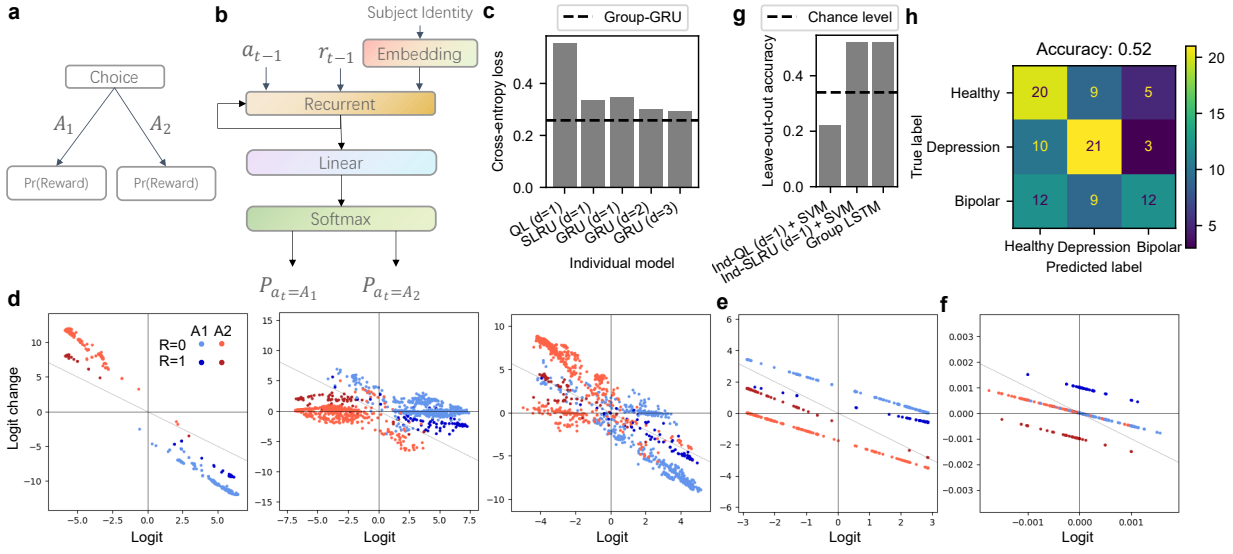
Given the superior accuracy of RNNs over classical cognitive models, we next aimed to identify cognitive strategies these RNNs learned, by applying the dynamical systems approach developed by Ji-An et al. (2023). This approach analyzes temporal changes in state variables (i.e., neuronal activations) 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 $\Pr_t(A_i)$ represents the probability of selecting action A_i. The logit-change $\delta L(t)$ is defined as $L(t+1) - L(t)$, the difference between logits of two consecutive trials. For each subject, we plotted the logit-change against logit for each trial provided by the Group-GRU, grouped by the four input conditions (2 actions by 2 rewards). We observed a remarkable degree of individual differences in logit patterns, suggesting that networks trained on group data do not emulate a singular group representative as hypothesized in Dezfouli et al. (2019). Instead, they simulate diverse strategies, tailoring their predictions to each subject's identity and distinct history.

We outline several representative strategies manifested by subjects (each can be validated when inspecting raw choice data) . **(i) Switching strategy** (Fig. 1d left). This is evidenced by the logit-change δL lying above the diagonal line ($\delta L = -L$) for red A_2 points with $L < 0$ and below the diagonal line for blue A_1 points 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, this subject frequently switches the action preference in each trial, explaining the oscillatory behavior previously observed in trained RNNs (Dezfouli et al. 2019), i.e., initial action oscillations trigger following action switches. **(ii) Inverted-reward strategy** (Fig. 1d middle). In classical model-free RL models (e.g., Q-learning), a reward following A_1 leads to a positive logit change (preferring A_1 more), while no reward following A_1 leads to a negative logit change (Ji-An et al. 2023; see Fig. 1f where the dark blue $[A_1\ R = 1]$ trials lie above light blue $[A_1\ R = 0]$ trials). Strikingly, RNNs discovered that, in most sub-

jects, a reward following A_1 usually leads to a negative logit change (dark blue points below the x-axis), decreasing the preference for A_1, suggesting a tendency to switch. The lack of reward following 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, not 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 (Dezfouli et al. 2019). **(iii) Alternations of strategies (i) and (ii)** (Fig. 1d right), i.e., using a strategy for several trials and then alternating to another strategy. Similar alternation phenomena were reported in rodents (Ashwood et al. 2022), but not in humans yet.

Next, we aimed to determine if a subject's logit patterns can predict their healthy and pathological status, a core objective for computational psychiatry. We first trained one-dimensional RNNs with switching linear recurrent unit (SLRU; Ji-An et al. 2023) on individual choice data (Ind-SLRU, $d = 1$), acting as a first-order approximation of a one-dimensional compression of these strategies. The hidden state h_t ($t > 0$) is updated by $h_t = W^{(x_{t-1})}h_{t-1} + b^{(x_{t-1})}$, where $W^{(x_{t-1})}$ and $b^{(x_{t-1})}$ are the weight matrices and biases determined by the input x_{t-1}. Importantly, these weights and biases are interpretable and directly related to the learning rate α ($W = 1 - \alpha$) and the reward r ($b = \alpha r$) in the RL models, respectively. We found that these Ind-SLRUs achieve an average loss of 0.337 (Fig. 1c), suggesting they provide a reasonably well one-dimensional summarization of behavior (comparing Fig. 1e to Fig. 1d middle for strategy (ii)). This significantly outperforms the one-dimensional QL model (i.e., $V(A_1, t+1) = (1-\alpha)V(A_1, t) + \alpha r \cdot \mathrm{Sgn}(a_t = A_1)$) fitted to individual choices (Ind-QL, $d = 1$, Fig. 1c), which cannot effectively summarize these underlying strategies (comparing Fig. 1f to Fig. 1d middle for strategy (ii)). We then extracted interpretable weights and biases from Ind-SLRUs as features to predict diagnostic labels. We found that the linear-kernel support vector machine (SVM) using Ind-SLRU features achieves an overall 52% accuracy (see Fig. 1g, h), substantially better than the SVM using Ind-QL features (i.e., learning rate and inverse temperature). Our approach differs from the original study that also achieved a 52% accuracy (Dezfouli et al. 2019), where authors trained separate LSTMs for each diagnostic group and assigned the label to the leave-one-out subject based on the Group-LSTM with the largest likelihood. Our results indicate that the diagnostic accuracy using interpretable SLRU features is comparable to that using larger, black-box RNNs.

To summarize, we harnessed the flexibility of tiny RNNs and dynamical-systems interpretability to characterize decision making in healthy and mentally ill individuals. We identified several cognitive strategies previously overlooked by classical cognitive modeling and an effective diagnostic procedure for computational psychiatry.



(a) Task structure. (b) RNN structure with a subject embedding layer. (c) Predictive performance (lower is better) for individual models (with different dimensionalities d) and the Group-GRU model. (d) Logit analysis of the group-GRU for strategies used by three example subjects. (Left) strategy (i). (Middle) strategy (ii). (Right) strategy (iii). (e) Logit analysis of the Ind-SLRU ($d = 1$) for strategy (ii). (f) Logit analysis of the Ind-QL ($d = 1$) for strategy (ii). (g) The diagnostic accuracy evaluated with leave-one-subject-out. (h) The confusion matrix for Ind-SLRU ($d = 1$) + SVM.