bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Models that learn how humans learn: the case of depression and bipolar disorders

Amir Dezfouli[1], Kristi Griffiths[2], Fabio Ramos[3], Peter Dayan[4†], Bernard W. Balleine[1†*]

1 School of Psychology, UNSW, Sydney, Australia
2 Westmead Institute for Medical Research, University of Sydney, Sydney, Australia
3 School of Information Technologies, University of Sydney, Sydney, Australia
4 Gatsby Computational Neuroscience Unit, UCL, London, UK
***bernard.balleine@unsw.edu.au**
[†]co-senior authors

Abstract

Computational models of learning and decision-making processes in the brain play an important role in many domains. Such models typically have a constrained structure and make specific assumptions about the underlying human learning processes; these may make them underfit observed behaviours. Here we suggest an alternative method based on learning-to-learn approaches, using recurrent neural networks (RNNs) as a flexible family of models that have sufficient capacity to represent the complex learning and decision-making strategies used by humans. In this approach, an RNN is trained to predict the next action that a subject will take in a decision-making task, and in this way, learns to imitate the processes underlying subjects' choices and their learning abilities. We demonstrate the benefits of this approach with a new dataset containing behaviour of uni-polar depression (n=34), bipolar (n=33) and control (n=34) participants in a two-armed bandit task. The results indicate that the new approach is better than baseline reinforcement-learning methods in terms of overall performance and its capacity to predict subjects' choices. We show that the model can be interpreted using off-policy simulations, and thereby provide a novel clustering of subjects' learning processes – something that often eludes traditional approaches to modelling and behavioural analysis.

Introduction

A computational model of learning in decision-making can be regarded as a mathematical function that inputs past experiences (such as chosen actions and the rewards that result), and outputs predictions of the actions that will be taken in the future (e.g., Busemeyer and Stout, 2002; Dezfouli et al., 2007; Montague et al., 2012; Daw et al., 2006). Typically, experimenters specify a set of assumptions that define and constrain the general structure and form of a whole class of computational models, leaving free a set of parameters that are estimated from the data. For example, in value-based decision-making, a common assumption is that subjects' choices are determined in a noisy manner by learned action values (often called Q values; Watkins, 1989), which are updated given experience of rewards. This model has two free parameters: the level of noise and the learning rate governing updating.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Despite the flexibility afforded by the free parameters, a class of computational models can only capture learning processes that fall within the boundaries of the assumptions embedded in its structure. If the actual learning and choice method used by the subjects is more complex or otherwise different from that implied by the structure of the model, then it will under- and/or otherwise mis-fit the data. For example, if the model assumes that subjects are using a single learning-rate parameter to update the values of actions, but in reality rewards and punishments are modulated by two different learning-rates, then the model will fail to provide a complete representation of the learning processes (e.g., Piray et al., 2014). Because of this, in practice, the process of computational modelling typically involves various forms of analysis in order to confirm that the assumptions about the behaviour are correct. If they are found wanting, then different assumptions must be made, leading to different models that must be selected between to find the one that misfits least. This iterative process has become standard scientific practice for model development, and has been influential in domains such as cognitive modelling, computational psychiatry (e.g., Busemeyer and Stout, 2002; Dezfouli et al., 2007; Montague et al., 2012), and model-based analysis of neural data (e.g., Daw et al., 2006).

By contrast, we consider an alternative approach to modelling involving minimal assumptions about the underlying learning processes used by subjects. Instead, these are captured by a very flexible class of models on the basis of observing the information the subjects see and the choices they make, and predicting the latter. This process is known as *learning-to-learn* (Hochreiter et al., 2001; Wang et al., 2016; Duan et al., 2016; Weinstein and Botvinick, 2017).[1] Since the models are flexible, they can come automatically to characterize the major behavioural trends exhibited by the subjects, without requiring tweaking and engineering explicitly based on behavioural analysis of data. This approach is particularly useful when major trends in the data are not apparent in behavioural summary statistics. Even if such trends are visible, it might be complicated to create compact models that encompass them adequately.

In particular, we consider as our flexible class recurrent neural networks (RNNs), which are known to have sufficient capacity to represent a wide range of learning processes used by humans (and other animals). Learning to learn involves adjusting the weights in the networks so they can predict the choices that subjects will make as those subjects themselves learn. Once the weights have been trained, they are frozen, and the model is simulated in the actual learning task to assess its predictive capacity and to gain insights into human behaviour.

To illustrate and evaluate this approach, we focus on a relatively simple decision-making task in which subjects had a choice between two key presses that were rewarded probabilistically (a two-arm bandit task). Data from three groups were collected: healthy subjects, and patients with depression and bipolar disorders. The results showed that the new method was able to learn subjects' decision-making strategies more accurately than baseline models. Furthermore, we show that off-policy simulations of the model help visualise, and thus uncover, the properties of the learning process behind subjects' actions. We show that these were inconsistent with the assumptions made by baseline reinforcement-learning treatments. Finally, we show how the method can be applied to predict diagnostic labels for different patient populations.

[1] Albeit more commonly the learner is a human facing a series of learning tasks, rather than a computer model trying to copy the human on a single task.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and methods

Participants

34 uni-polar depression (DEPRESSION), 33 bipolar (BIPOLAR) and 34 control (HEALTHY) participants (age, gender, IQ and education matched) were recruited from outpatient mental health clinics at the Brain and Mind Research Institute, Sydney, and the surrounding community. Participants were aged between 16 and 33 years. Exclusion criteria for both clinical and control groups were history of neurological disease (e.g. head trauma, epilepsy), medical illness known to impact cognitive and brain function (e.g. cancer), intellectual and/or developmental disability and insufficient English for neuropsychological assessment. Controls were screened for psychopathology by a research psychologist via clinical interview. Patients were tested under 'treatment-as-usual' conditions, and at the time of assessment, 77% of depressed and 85% of bipolar patients were taking medications (see Table 1 for breakdown of medication use). The study was approved by the University of Sydney ethics committee. Participants gave informed consent prior to participation in the study.

Demographics and clinical characteristics of the sample are presented in Table 1. Levene's test indicated unequal variances for the HDRS (Hamilton Depression Rating Scale; Hamilton, 1960), YMRS (Young Mania Rating Scale; Young et al., 1978), SOFAS (Social and Occupational Functional Scale; Goldman et al., 1992) and age, thus Welch's statistic was used for these variables. A one-way ANOVA revealed no differences between groups in age $[F(2, 98) = 2.48,\ p = 0.09]$, education $[F(2, 98) = 1.76,\ p = 0.18]$, IQ $[F(2, 94) = 0.47,\ p = 0.62]$ or gender $(\chi^2 = 2.66,\ p = 0.27)$. There were differences in HDRS $[F(2, 49.21) = 64.21,\ p < 0.001]$, YMRS $[F(2, 43.71) = 12.57,\ p < 0.001]$, and SOFAS $[F(2, 41.61) = 169.66,\ p < 0.001]$. Bonferroni post-hoc comparisons revealed higher depression scores in DEPRESSION group compared to BIPOLAR and HEALTHY groups, and higher depression in BIPOLAR group compared to HEALTHY group. Mania scores were significantly higher in BIPOLAR group compared to HEALTHY group. Both patient groups had significantly reduced SOFAS scores compared to HEALTHY group, but did not differ from one another. Age of mental illness onset was younger in DEPRESSION group compared to BIPOLAR group $[t(56) = -2.14,\ p = 0.04]$, however duration of illness did not significantly differ between groups $[t(56) = 1.25,\ p = 0.22]$. There were no differences between groups in pre-test hunger $[F(2, 79) = 0.54,\ p = 0.59]$ or average snack rating $[F(2, 79) = 2.53,\ p = 0.09]$.

Task

The instrumental learning task (Figure 1) involved participants choosing between pressing the left or right button in order to earn food rewards (an M&M chocolate or a BBQ flavoured cracker). We refer to these two key presses as L and R for left and right button presses respectively. Fourteen HEALTHY participants (41.2% of the group) and 13 BIPOLAR participants (36.7% of the group) completed the task in an fMRI setting, using a 2 button Lumina response box. The remaining HEALTHY and BIPOLAR participants, and all DEPRESSION participants, completed the task on a computer with a keyboard, where the "Z" and "?" keys were given as L and R. Although the performance of subjects was overall higher in the fMRI settings $[\beta = 0.050,\ \text{SE}= 0.024,\ p = 0.041]$[2], the mode of task completion (in fMRI setting vs on a computer) had no significant effect on how choices were adjusted on a trial-by-trial basis, and

[2] The intercept term was random-effect at the group level (HEALTHY or BIPOLAR), and the mode of task completion (in fMRI settings vs on a computer) was the fixed-effect; the probability of staying on the same action was the dependent variable; see section Statistical analysis for details.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Demographic and clinical characteristics of participants. Means(SD).
HDRS: Hamilton Depression Rating Scale;
YMRS: Young Mania Rating Scale;
SOFAS: Social and Occupational Functioning Scale;
a: DEPRESSION greater than HEALTHY and BIPOLAR, $p < 0.05$.
b: BIPOLAR greater than HEALTHY, $p < 0.05$.
c: HEALTHY greater than DEPRESSION and BIPOLAR, $p < 0.05$.
d: DEPRESSION slower than HEALTHY and BIPOLAR, $p < 0.05$.

	HEALTHY (n=34)	DEPRESSION (n=34)	BIPOLAR (n=33)
Demographics			
Gender (M:F)	15:19	15:19	9:24
Age in years	23.6 (4.3)	21.6 (2.5)	23.1 (4.4)
Predicted IQ	107.3 (7.5)	105.5 (7.9)	106.0 (7.4)
Eduation	14.3 (3.0)	13.3(1.9)	13.3 (2.4)
Symptoms and History			
Age of onset (years)	-	14.4 (3.8)	15.9 (4.7)
Duration of illness (years)	-	7.7 (4.3)	6.4 (3.3)
HDRS	1.5(2.0)	14.1 (7.2)^a	8.9 (6.5)
YMRS	0.1 (0.4)	2.5 (5.4)	4.6 (5.8)^b
SOFAS	91.0 (3.5)^c	63.8 (9.2)	65.7 (13.7)
Medication			
Medicated	-	77%	85%
Anti-depressants	-	71%	41%
Mood stabilizers/Anti-convulsants	-	9%	73%
Lithium	-	0%	18%
Anti-psychotics	-	18%	33%
Anxiolytics	-	0%	3%
Motivation measures			
Hunger	6.5 (1.7)	6.0 (2.1)	6.0 (2.4)
Reward Pleasantness	3.1 (1.3)	2.0 (2.0)	2.6 (2.0)
Press Rate/Sec	0.93 (0.3)	0.71 (0.2)^d	0.98 (0.3)

Duration of illness indicates time since patient first experienced mental health problems, not time since diagnosis.

therefore the data from both groups were combined (how subjects completed the task had no significant effect on the probability of staying on the same action neither after earning a reward [$\beta = 0.041$, SE$= 0.054$, $p = 0.45$][3], nor after no reward [$\beta = 0.030$, SE$= 0.062$, $p = 0.627$]).

During each block, one action was always associated with a higher probability of reward than the other. Across blocks, the action with the higher reward probability switched identities (left or right), and the probabilities varied (taking one of the values 0.25, 0.125, 0.08). The probability of reward on the other action always remained at 0.05. Participants were instructed to earn as many points as possible, as they would be given the concomitant number of M&Ms or BBQ flavoured crackers at the end of the session. During each a non-rewarded response, a grey circle appeared in the centre of the screen for 250ms, whereas during a rewarded response, the key turned green and an image of the food reward

[3] The intercept term was random-effect at the group level (HEALTHY or BIPOLAR), and the mode of task completion (in fMRI settings vs on a computer) was the fixed-effect; the probability of selecting the better key was the dependent variable; see section Statistical analysis for details.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

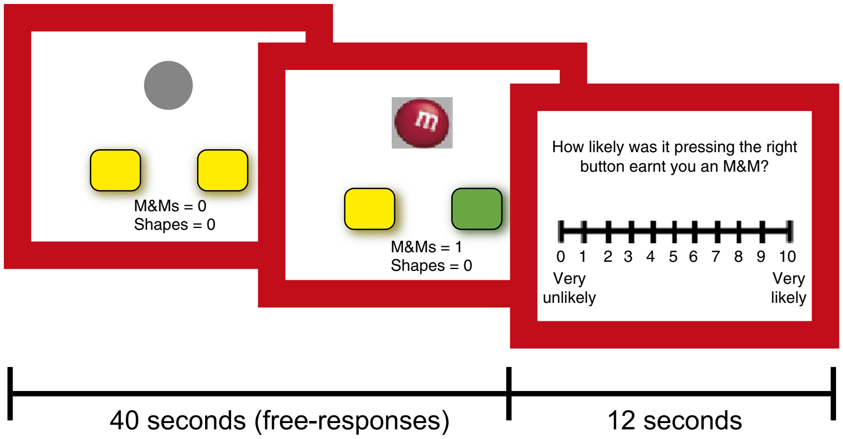

Figure 1. Structure of the decision-making task. Before the choice, no stimulus indicated which button was more likely to lead to reward. When the participant made a choice, the button chosen was highlighted (green) and on rewarded trials the reward stimulus was presented for 500ms duration. After each block of trials, the participant rated how causal each button was in earning rewards.

appeared in the centre of the screen for 500ms. A tally of accumulated winnings remained on the bottom of the screen for the duration of the task. Responding was self-paced during the 12 blocks, each 40-s in length. At the end of each block, participants were asked to judge, on a 10-point scale, how likely it was that pressing each button earned them the reward on the previous trial.

The task began with a 0.25 contingency practice block, a hunger rating (0 to 10) and a pleasantness rating for each food outcome (-5 to +5). The data from hunger ratings and subjective ratings (at the end of each block) was missing for some subjects and it was not used in the analysis.

Computational models

Notation

The set of available actions is denoted by $\mathcal{A}$. Here $\mathcal{A} = \{\mathrm{L}, \mathrm{R}\}$, with L and R referring to left and right key presses respectively. A set of subjects is denoted by $\mathcal{S}$, and the total number of trials completed by subject $s \in \mathcal{S}$ over the whole task (all blocks) is denoted by $\mathcal{T}_s$. a_t^s denotes the action taken by subject s at trial t. The reward earned at trial t is denoted by r_t, and we use a_t to refer to an action taken at time t, either by the subjects or the models (in simulations).

Recurrent neural network model (RNN)

The architecture used is based on recurrent neural network model (RNN) and is depicted in Figure 2. The model is composed of an LSTM layer (Long short-term memory; Hochreiter and Schmidhuber, 1997) and an output softmax layer with two nodes (since there are two actions in the task). The inputs to the LSTM layer are the previous action (a_{t-1} coded using one-hot transformation) and the reward received after taking action ($r_{t-1} \in \{0, 1\}$). The outputs of the softmax are probabilities of selecting each action, which are denoted by $\pi_t(a; \textsc{rnn})$ for action $a \in \mathcal{A}$ at trial t.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the learning-to-learn phase, the aim is to train weights in the network so that the model learns to predict subjects' actions given their past observations (i.e., learns how *they* learn). For this purpose, the objective function for optimising weights in the network (denoted by Θ) for subject set $\mathcal{S}$ is,

$$\mathcal{L}(\Theta;\text{RNN}) = \sum_{s\in\mathcal{S}} \sum_{t=1\ldots\mathcal{T}_s} \log \pi_t(a_t^s;\text{RNN}), \tag{1}$$

where a_t^s is the action selected by subjects s at trial t, and $\pi_t(.;\text{RNN})$ is the probability that model assigns to each action. Note that the policy is conditioned on the previous actions and rewards in each block of training, which are not shown in notations for simplicity.

Models were trained using maximum-likelihood (ML) estimation method,

$$\Theta_{\text{RNN}}^{\text{ML}} = \arg\max_{\Theta} \mathcal{L}(\Theta;\text{RNN}), \tag{2}$$

where Θ is a vector containing free-parameters of the model (in both LSTM and softmax layers). The models were implemented in TensorFlow (Abadi et al., 2016) and optimized using Adam optimizer (Kingma and Ba, 2014). Note that Θ was estimated for each group of subjects separately. Networks with different numbers N_c of LSTM cells ($N_c \in \{5, 10, 20\}$) were considered, and the best model was selected using leave-one-out cross-validation (see below). Early stopping was used for regularization and the optimal number of training iterations was selected using leave-one-out cross-validation.

The total number of free parameters (in both the LSTM layer and softmax layer) were 190, 580, and 1960 for the networks with 5, 10, and 20 LSTM cells, respectively. In order to control for the effect of initialization of network weights on the final results, a single random network of each size (5, 10, 20) was generated, and was used to initialize the weights in the network.

After the learning-to-learn phase, the weights in the network were frozen and the trained model was used for three purposes: (i) cross-validation (see below), (ii) on-policy simulations and (iii) off-policy simulations. For cross-validation, the previous actions of the test subject(s) and the rewards experienced by the subject(s) were fed into the model, but unlike the learning-to-learn phase, the weights were not changing and we only recorded the prediction of the model about the next action. Note that even though the weights in the network were fixed, the output of the network changed from trial to trial due to the recurrent nature of these networks. Also, due to the small sample size we used the same set of subjects for testing the model and for the validation of model hyper-parameters (N_c and number of optimization iterations).

Other than being used for calculating cross-validation statistics, trained models were used for on-policy and off-policy simulations (with frozen weights). In the on-policy simulations, the model received its own actions and earned rewards as the inputs (instead of receiving the action selected by the subjects). In the off-policy simulations, the set of actions and rewards that the model received was fixed and predetermined. The details of these simulations are reported in the Results section.

Baseline methods

We used three baseline methods, QL, QLP and GQL, which are variants and generalizations of Q-learning (Watkins, 1989).

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

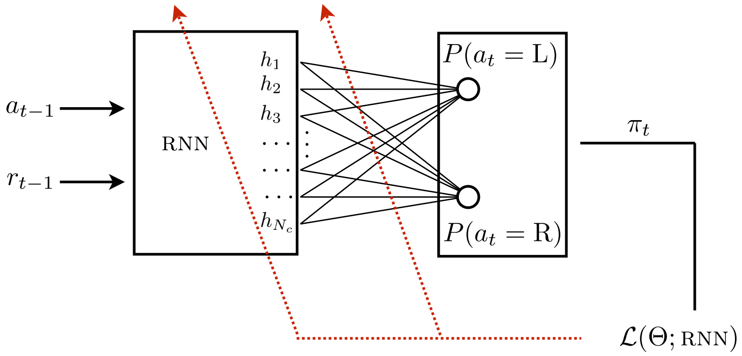

Figure 2. Structure of the RNN model. The model has a LSTM layer which receives previous action and reward as inputs, and is connected to a softmax layer which outputs the probability of selecting each action in the next trial (policy). Maximum likelihood estimate method is used to train the weights in LSTM and softmax layers. h_i is the output of LSTM cell i in the LSTM layer. N_c is the number of cells in the LSTM layer.

QL model. After taking action a_{t-1} at time $t-1$, the value of the action, denoted by $Q_t(a_{t-1})$, is updated as follows,

$$Q_t(a_{t-1}) = (1-\phi)Q_{t-1}(a_{t-1}) + \phi r_{t-1}, \tag{3}$$

where ϕ is the learning-rate and r_{t-1} is the reward received after taking the action. Given the action values, the probability of taking action $a \in \{\mathrm{L}, \mathrm{R}\}$ in trial t is:

$$\pi_t(a; \textsc{ql}) = \frac{e^{\beta Q_t(a)}}{\sum_{A' \in \mathcal{A}} e^{\beta Q_t(a')}},$$

where $\beta > 0$ is a free-parameter and controls the contribution of values to the choices (balance between 148
exploration and exploitation). The free-parameters of this variant are ϕ and β. Note that the probability 149
that models predict for each action at trial t, is necessarily based on the data *before* observing the action 150
and reward at trial t. Further, since there are only two actions, we can write 151
$\pi_t(\mathrm{L}; \textsc{ql}) = 1 - \pi_t(\mathrm{R}; \textsc{ql}) = \sigma(\beta(Q_t(\mathrm{L}) - Q_t(\mathrm{R})))$ where $\sigma(\cdot)$ is the standard logistic sigmoid. 152

QLP model. This model is inspired by the fact that humans and other animals have a tendency to stick with the same action on multiple trials (i.e., perseverate), or sometimes to alternate between the actions (independent of the reward effects; Lau and Glimcher, 2005). We therefore call this model QLP, for Q-learning with perseveration. In it, action values are updated according to equation 3 similar to QL model, but the probability of selecting actions is,

$$\pi_t(a; \textsc{qlp}) = \frac{e^{\beta Q_t(a) + k_t(a)}}{\sum_{a' \in \mathcal{A}} e^{\beta Q_t(a') + k_t(a')}},$$

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where,

$$k_t(a) = \begin{cases} \kappa & \text{if } a = a_{t-1} \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Therefore, there is a tendency for selecting the same action again in the next trial (if $\kappa > 0$) or switching to the other action (if $\kappa < 0$), and, in the specific case that $\kappa = 0$, the QLP model reduces to QL. Free-parameters are ϕ, β, κ.

GQL model. As we will show in the results section, neither QL nor QLP fit the behaviour of the subjects in the task. As such, we aimed to develop a baseline model which could at least capture high-level behavioural trends, and we built a generalised Q-learning model, GQL, to compare with RNN. In this variant, instead of learning a single action value for each action, the model learns N different values for each action, where the difference between the values learned for each action is that they are updated using different learning-rates. The action values for action a are denoted by $\mathbf{Q}(a)$, which is a vector of size N, and the corresponding learning-rates are denoted by vector Φ of size N ($\mathbf{0} \preceq \Phi \preceq \mathbf{1}$). Based on this, the value of action a_{t-1} at trial $t-1$ is updated as follows,

$$\mathbf{Q}_t(a_{t-1}) = (\mathbf{1} - \Phi) \odot \mathbf{Q}_{t-1}(a_{t-1}) + r_{t-1}\Phi, \tag{5}$$

where $\odot$ represents element-wise Hadamard product. For example, if $N = 2$, and $\Phi = [0.1, 0.05]$, then it means that the model learns two different values for each action (L, R actions) and one of the values will be updated using learning-rate 0.1 and the other one is updated using learning-rate 0.05. In the specific case that $N = 1$, the above equation reduces to equation 3 used in QL and QLP models, in which only a single value is learned for each action.

In the QLP model, the current action is affected by the last taken action (perseveration). This property is generalised in GQL model by learning the history of previously taken actions, instead of just the last action. These action histories are denoted by $\mathbf{H}(a)$ for action a. $\mathbf{H}(a)$ is a vector of size N, and each entry of this vector tracks the tendency of taking action a in the past, i.e., if an element of $\mathbf{H}(a)$ is close to one it means that action a was taken frequently in the past and being close to zero implies that the action was taken rarely. Similar to action values, for each action, N different histories are tracked, each of which is modulated by a separate learning-rate. Learning-rates are represented in vector Ψ of size N ($\mathbf{0} \preceq \Psi \preceq \mathbf{1}$). Assuming that action a_{t-1} was taken at trial $t-1$, $\mathbf{H}(a)$ updates as follows,

$$\mathbf{H}_t(a) = \begin{cases} (\mathbf{1} - \Psi) \odot \mathbf{H}_{t-1}(a) + \Psi & \text{if } a = a_{t-1} \\ (\mathbf{1} - \Psi) \odot \mathbf{H}_{t-1}(a) & \text{otherwise} \end{cases}. \tag{6}$$

Intuitively, according to the above equation, if action a was taken on a trial, $\mathbf{H}(a)$ increases (the amount of increase depends on the learning-rate of each entry), and for the rest of the actions, $\mathbf{H}$(other actions) will decrease (again the amount of decrement is modulated by the learning rates). For example, if $N = 2$, and $\Psi = [0.1, 0.05]$, it means that for each action two choice tendencies will be learned, one of which is updated by rate 0.1 and the other one by rate 0.05.

Having learned $\mathbf{Q}(a)$ and $\mathbf{H}(a)$ for each action, the next question is how these two combine to guide

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

choices. Q-learning models assume that the contribution of values to choices is modulated by parameter β. Here, since the model learns multiple values for each action, we assume that each value is weighted by a separate parameter, denoted by vector $\mathbf{B}$ of size N. Similarly, in the QLP model the contribution of perseveration to choices is controlled by parameter κ, and here we assume that parameter $\mathbf{K}$ modulates the contribution of previous actions to the current choice. Based on this, the probability of taking action a at trial t is,

$$\pi'_t(a;\textsc{gql}) = \frac{e^{\mathbf{B}\cdot\mathbf{Q}_t(a)+\mathbf{K}\cdot\mathbf{H}_t(a)}}{\sum_{a'\in\mathcal{A}} e^{\mathbf{B}\cdot\mathbf{Q}_t(a')+\mathbf{K}\cdot\mathbf{H}_t(a')}},$$

where "$\cdot$" operator refers to inner product. Here, we also add an extra flexibility to the model by allowing values to interact with history of previous actions for influencing choices. For example, if $N = 2$, we allow the two learned values for each action to interact with the two learned action histories of each action, leading to four interaction terms, and the contribution of each interaction term to choices is determined by matrix $\mathbf{C}$ of size $N \times N$ ($N = 2$ in this example),

$$\pi_t(a;\textsc{gql}) = \frac{e^{\mathbf{B}\cdot\mathbf{Q}_t(a)+\mathbf{K}\cdot\mathbf{H}_t(a)+\mathbf{H}_t(a)\cdot\mathbf{C}\cdot\mathbf{Q}_t(a)}}{\sum_{a'\in\mathcal{A}} e^{\mathbf{B}\cdot\mathbf{Q}_t(a')+\mathbf{K}\cdot\mathbf{H}_t(a')+\mathbf{H}_t(a')\cdot\mathbf{C}\cdot\mathbf{Q}_t(a')}}, \tag{7}$$

The free-parameters of this model are Φ, Ψ, $\mathbf{B}$, $\mathbf{K}$, and $\mathbf{C}$. In this paper we use models with $N = 1, 2, 10$, which have 5, 12 and 140 free parameters respectively. We used $N = 2$ for the results reported in the main text, since this model setting was able to capture several behavioural trends while still being interpretable. The results using $N = 1, 10$ are reported in the supplementary materials to illustrate the models' capabilities in extreme cases.

Objective function. The objective function for optimising the models was the same as the one chosen for RNN,

$$\mathcal{L}(\Theta;\mathcal{M}) = \sum_{s\in\mathcal{S}}\sum_{t=1\ldots\mathcal{T}_s} \log \pi_t(a^s_t;\mathcal{M}), \mathcal{M} \in \{\textsc{ql}, \textsc{qlp}, \textsc{gql}\}, \tag{8}$$

where as mentioned before, a^s_t is the action selected by subject s at trial t, and $\pi_t(.;\mathcal{M})$ is the probability that model $\mathcal{M}$ assigns to each action. Models were trained using maximum-likelihood estimation method,

$$\Theta^{\text{ML}}_{\mathcal{M}} = \arg\max_{\Theta} \mathcal{L}(\Theta;\mathcal{M}), \tag{9}$$

where Θ is a vector containing the free-parameters of the models. Optimizations for all models were performed using Adam optimizer (Kingma and Ba, 2014), and using automatic differentiation method provided in TensorFlow (Abadi et al., 2016). The free-parameters with the limited support (ϕ, β, Φ, Ψ) were transformed to satisfy the constraints.

Performance measures

Two different measures were used for quantifying the predictive accuracy of the models. The first measure is the average log-probability of the models' prediction for the actions taken by subjects. For a

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

group of subjects denoted by $\mathcal{S}$, we define negative log-probability (NLP) as follows:

$$\text{NLP} = -\frac{\sum_{s\in\mathcal{S}}\sum_{t=1\ldots\mathcal{T}_s}\log\pi_t(a_t^s;\mathcal{M})}{\sum_{s\in\mathcal{S}}\mathcal{T}_s}, \mathcal{M}\in\{\text{RNN}, \text{GQL}, \text{QL}, \text{QLP}\}. \tag{10}$$

The other measure is the percentage of actions predicted correctly,

$$\%\text{correct} = \frac{\sum_{s\in\mathcal{S}}\sum_{t=1\ldots\mathcal{T}_s}[\![\arg\max_a \pi_t(a;\mathcal{M}) = a_t^s]\!]}{\sum_{s\in\mathcal{S}}\mathcal{T}_s}, \tag{11}$$

where $[\![.]\!]$ denotes the indicator function. Unlike, '%correct', NLP takes the probabilities of predictions into account instead of making binary predictions for the next action. In this way, if the models are certain about wrong predictions NLP performance gets penalized, and it gets credit if the models are certain about a correct prediction.

Model selection

Leave-one-out cross-validation was used for comparing different models. At each round, one of the subjects was held out and the model was trained using the rest of the subjects; the trained model was then used for making predictions for the held out subject. The held out subject was rotated in the each group, yielding 34, 34, 33 prediction accuracy measures in HEALTHY, DEPRESSION, and BIPOLAR groups respectively.

Statistical analysis

For the analysis we performed hierarchical linear mixed-effects regression using the lme4 package in R (Bates et al., 2015) and obtained p-values for regression coefficients using the lmerTest package (Kuznetsova et al., 2016). For each test we report parameter estimate (β), standard error (SE), and p-value.

Results

We first focus on the high-level evaluation of the new approach in terms of making predictions about subjects' actions and diagnostic labels. Then, in the following sections we focus on behavioural analysis of the data using the RNN.

Prediction analysis

Model settings. For the RNN model, leave-one-out cross-validation was used to determine the number of cells and optimisation iterations required for the RNN model to achieve the highest prediction accuracy. The results are shown in Figure S1, which shows that the lowest mean negative log-probability (NLP) is achieved by 10 cells in the LSMT layer and after 1100, 1200 optimisation iterations for HEALTHY and DEPRESSION groups respectively. For BIPOLAR group the best NLP was achieved by 20 cells and 400 optimisation iterations. These settings were used in the next steps for making predictions and simulations. For the case of the GQL model, we used $N = 2$, which implies that for each action two different values and action histories were tracked by the model.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

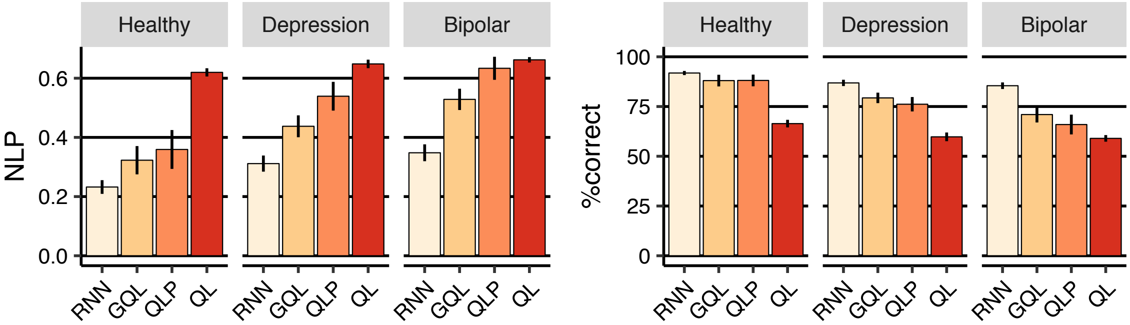

Figure 3. Cross-validation results. **(Left-panel)** NLP (negative log-probability) averaged across leave-one-out cross-validation folds. Lower values are better. **(Right-panel)** Percentage of actions predicted correctly averaged over cross-validation folds. Error-bars represent 1SEM.

Action prediction. Here the aim was to quantify how well the models predict the actions chosen by the subjects. We used Leave-one-out cross-validation (as above) and calculated prediction accuracy measures for the held out subjects. The results are reported in Figure 3. Left-panel in the figure shows prediction accuracy in terms of NLP (averaged over leave one-out cross-validation folds; lower values are better) and the right-panel shows the percentage of actions predicted correctly (higher values are better). Focusing on NLP measures, which unlike '%correct' take the certainty of predictions into account, we make two observations. Firstly, among the baseline models, GQL provided the highest performance in terms of NLP in all the three groups, which was statistically significant in DEPRESSION and BIPOLAR groups (HEALTHY [$\beta = -0.036$, SE$= 0.020$, $p = 0.086$][4], DEPRESSION [$\beta = -0.101$, SE$= 0.024$, $p < 0.001$], BIPOLAR [$\beta = -0.105$, SE$= 0.019$, $p < 0.001$]). Secondly, across all groups, RNN provided the highest mean performance (HEALTHY [$\beta = 0.090$, SE$= 0.040$, $p = 0.030$], DEPRESSION [$\beta = 0.126$, SE$= 0.021$, $p < 0.001$], BIPOLAR [$\beta = 0.180$, SE$= 0.032$, $p < 0.001$]. Based on this, we conclude that RNN was more predictive than the baseline models. Indeed, we show in the next section that it captures some trends in the behaviour of the subjects that other models fail to capture.

Diagnostic label prediction. Next, we sought to evaluate the new approach in terms of predicting diagnostic labels using a leave-one-out cross-validation method. In each run, one of the subjects were held out, and a RNN model was fitted to the rest of the subject's group. This model, along with the versions of the same model fitted on all the subjects in each of the other two groups, were used to predict the diagnostic label for the held out subject. This prediction was based on which of the three models provided the best fit (lowest NLP) for that subject. The results are reported in Table 2. The baseline random performance is near 33%. As the table shows, the highest performance is achieved in the HEALTHY group, in which 64% of subjects are classified correctly. On the other hand, in DEPRESSION group there is a significant portion of subjects classified as HEALTHY. The overall correct classification rate of the model is 52%, while GQL achieved 50% accuracy (Table S1). We therefore conclude that although GQL was unable to accurately characterize behavioural trends in the data (as we will show below), the group differences that were captured by GQL were sufficient to guide diagnostic label predictions.

[4] The intercept term was random-effect at the cross-validation fold level; model (GQL =1, QLP/RNN =0) was the fixed-effect.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Prediction of diagnostic labels using RNN. Number of subjects for each true- and predicted-labels. The numbers inside parenthesis are the percentage of number subjects relative to the total number of subjects in each diagnostic group.

		predicted labels		
		HEALTHY	DEPRESSION	BIPOLAR
true labels	HEALTHY	22 (64%)	8 (23%)	4 (11%)
	DEPRESSION	13 (38%)	16 (47%)	5 (14%)
	BIPOLAR	9 (27%)	9 (27%)	15 (45%)

Behavioural analysis

Subject's performance

Figure 4 shows the probability of selecting the better key (the key with the higher reward probability). Results for subjects are shown by SUBJ in the graph. The probability of selecting the better key was significantly higher than the other key in all groups (HEALTHY [$\beta = 0.270$, SE$= 0.026$, $p < 0.001$][5], DEPRESSION [$\beta = 0.149$, SE$= 0.028$, $p < 0.001$], BIPOLAR [$\beta = 0.119$, SE$= 0.021$, $p < 0.001$]). Comparing HEALTHY and DEPRESSION groups, revealed that group by key interaction had a significant effect on the probability of selecting actions [$\beta = -0.120$, SE$= 0.038$, $p = 0.002$][6]. A similar effect was observed when comparing HEALTHY and BIPOLAR groups [$\beta = -0.150$, SE$= 0.034$, $p < 0.001$]. In summary, these results indicate that all groups were able to direct their actions toward the better choice, however DEPRESSION and BIPOLAR groups were less able to do so compared to the HEALTHY group.

Models' performance

The aim here was to use model simulations to gain insights into the subjects' learning processes in different groups. For each model, three different instances were trained using subjects' actions in each group (estimated parameters for GQL, QLP and QL models are shown Tables S4, S3 and S2 respectively). For the case of RNN, the number of cells, optimization iterations and model initialisation were based on the numbers obtained using cross-validation (see above). Negative log-likelihood for each model is reported in Table S5 (see Table S7 for the effect of the initialisation of the network on the negative log-likelihood of trained RNN. See Table S6 for negative log-likelihood when a separate model is fitted to each subject in the case of baseline models).

Models were simulated on-policy (i.e., actions were selected by the model according to the probabilities that the model assigned to each action on each trial) in the task conditions with the same probabilities and for the same number of trials that each subject completed in each block. The results of the simulations are shown in Figure 4 (RNN, GQL, QL, QLP). In the case of RNN, similar to the subjects' data, the probability of selecting the better key was significantly higher than the other key in all the three groups (HEALTHY [$\beta = 0.192$, SE$= 0.011$, $p < 0.001$], DEPRESSION [$\beta = 0.058$, SE$= 0.014$, $p < 0.001$], BIPOLAR [$\beta = 0.074$, SE$= 0.011$, $p < 0.001$]). Therefore, although the structure of RNN is initially unaware that the objective of the task is to collect rewards, its actions were directed toward the better key by following the strategy that it had learned from subjects' actions. A similar pattern was

[5] The intercept term was random-effect at the subject level; key (low reward probability=0, high reward probabilities=1) was the fixed-effect; dependent variable was the probability of selecting the key.

[6] The intercept term was the random-effect at the subject level; and key (low reward probability=0, high reward probabilities=1), groups (HEALTHY, DEPRESSION/BIPOLAR) and their interaction were fixed-effects; dependent variable was the probability of selecting the key.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

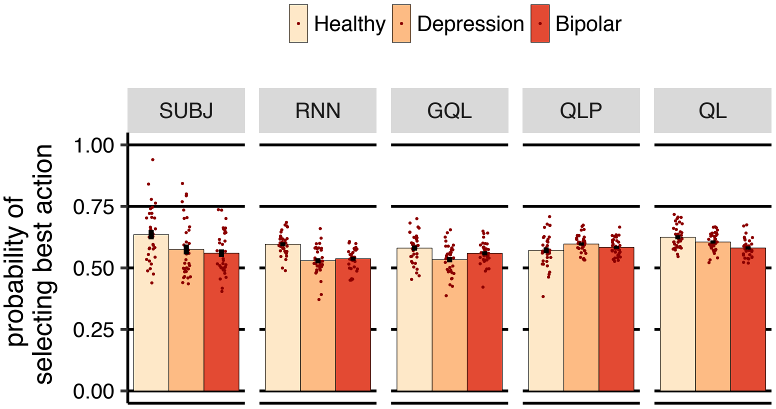

Figure 4. Probability of selecting the key with the higher reward probability (averaged over subjects). Each dot represents a subject and error-bars represent 1SEM.

observed for GQL, QLP and QL models, which is not surprising as the structure of these models includes value representations which can be used for reward maximization.

Immediate effect of reward on choices

Figure 5 shows the effect of earning a reward in the previous trial on the probability of staying on the same action in the next trial. For the subjects (SUBJ), earning a reward significantly *decreased* the probability of staying on the same action in HEALTHY and DEPRESSION groups, but not in BIPOLAR group (HEALTHY [$\beta = 0.112$, SE$= 0.019$, $p < 0.001$][7], DEPRESSION [$\beta = 0.111$, SE$= 0.029$, $p < 0.001$], BIPOLAR [$\beta = 0.030$, SE$= 0.035$, $p = 0.391$]). As the figure shows, the same pattern was observed in RNN (HEALTHY [$\beta = 0.082$, SE$= 0.006$, $p < 0.001$], DEPRESSION [$\beta = 0.089$, SE$= 0.013$, $p < 0.001$], BIPOLAR [$\beta = 0.001$, SE$= 0.010$, $p = 0.887$]), but stay probabilities had opposite directions in QL and QLP, i.e., the probability of staying on the same action was *higher* after earning reward (for the case of QLP; HEALTHY [$\beta = -0.028$, SE$= 0.004$, $p < 0.001$], DEPRESSION [$\beta = -0.039$, SE$= 0.006$, $p < 0.001$], BIPOLAR [$\beta = -0.054$, SE$= 0.007$, $p < 0.001$]), which differs from subjects' data.

This pattern is expected from baseline reinforcement-learning models, i.e., QL and QLP, as in these models, earning rewards increases the value of the taken action, which raises the probability of taking that action in the next trial. Indeed, such a learning process is embedded in the parametric forms of QL and QLP models, and cannot be reversed no matter what values are assigned to the free-parameters of these models. As such, we designed GQL as a baseline model with more relaxed assumptions and a higher capacity compared to QLP and QL, which enabled it to produce the same pattern as the subjects' choices, similar to RNN. Despite this, GQL provided a lower performance in terms of predicting subjects' choices compared to RNN, which shows there are behavioural trends that this model failed to represent, even though it was able to capture high-level behavioural statistics. Furthermore, it is not immediately clear how actions were directed toward the better key, and at the same time the probability of switching to the other action after earning rewards is higher, as it seems to imply actions will be diverted from the better key. In the next section, we aim to show how these two observations can be explained using off-policy model simulations.

[7] The intercept was random-effect at the subject level; whether reward was earned in the previous trial was the fixed-effect.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

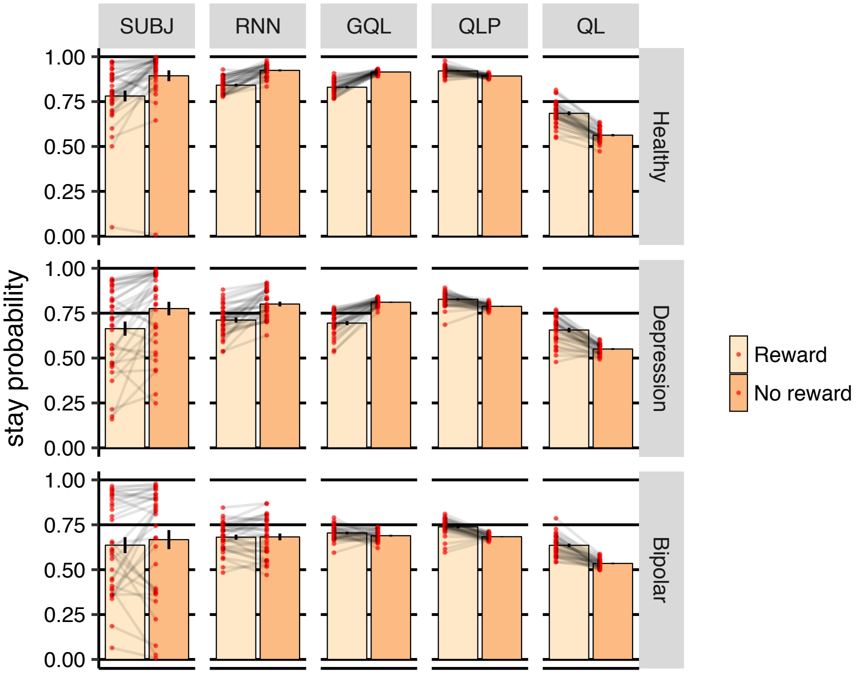

Figure 5. Probability of staying on the same action based on whether previous trial was rewarded (Reward) or no reward was earned (No reward), averaged over subjects. Each dot represents a subject and error-bars represent 1SEM.

Off-policy simulations 294

In this section we aim to use off-policy simulations of the models to uncover the learning and action-selection processes behind subjects' choices. Off-policy means that actions are not selected by the model in the simulation, but they are fixed and fed into the model, and the model is used only for making prediction about the next action. In this way we can control what inputs the model receives and thus examine how they affect predictions.

Simulations of the models (rows) are shown in Figure 6 for HEALTHY group, in which each panel shows a separate simulation across 30 trials (horizontal axis). For trials 1-10, the action that was fed to the model was R, and for trials 11-30 it was L (the action fed into the model at each trial is shown in the ribbons below each panel). The rewards associated with these trials varied among simulations (the columns) and are shown by black crosses (x) in the graphs (see Section S2 for more details on how simulation parameters were chosen).

Focusing on the RNN model, we can see that in the first 10 trials the predicted probability of taking R is higher than L; this then reverses in the next 20 trials. This shows that perseveration (i.e., sticking with the previously taken action) is an element in action selection, and is also consistent with the fact that the QLP model (which has a parameter for perseveration) performs better than the QL model in the cross-validation statistics (see Figure 3)[8]. We make four further sets of observations regarding how choices are affected by the history of previous rewards and actions.

[8] It is visible in Figure 4 that the probability of staying on an action is above 50%, irrespective of whether a reward was earned in the previous trial or not. This, however, does not provide any evidence for perseveration, as trials are not statistically independent. For example, in late training trials, a subject might have discovered which action returns more reward on average, and therefore stays on an action irrespective of reward, without necessarily relying on perseveration.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

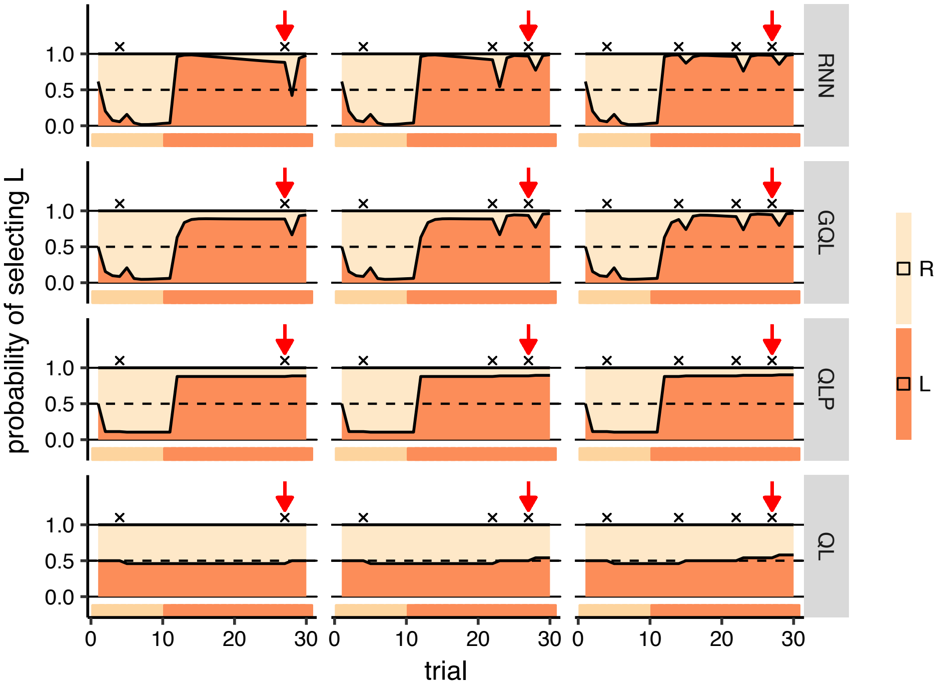

Figure 6. Off-policy simulations of all models for group HEALTHY. Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph. Note that the models' prediction for each trial is *before* seeing which action and reward was fed to the model on that trial.

The immediate effect of reward on choices. Focusing on RNN simulations in Figure 6, an observation is that earning a reward (shown by black crosses) causes a 'dip' in the probability of staying on an action, which shows the tendency to switch to the other action. This is consistent with the observation made in Figure 5 that the probability of switching increases after reward. We see a similar pattern in GQL, but in QL and QLP models the pattern is reversed, i.e., the probability of choosing an action increases after a reward due to the increases in action values (the effects are rather small for QLP and may not be clear for this model), which is again consistent with the observation in Figure 5. The reason that GQL is able to produce different predictions to that of QL and QLP is that in this model, the contribution of action values to choices can be negative, i.e., higher values can lead to lower a probability of staying on an action (see Section S1 for more explanation).

The effect of previous rewards on choices. The next observation is with respect to the effect of previous rewards on the probability of switching after a reward. First we focus on the RNN model and on the trials shown by red arrows in Figure 6. The red arrows point to the same trial number, but the number of rewards earned prior to the trial is different. As the figure shows, the probability of switching after reward is lower in the right-panel compared to the left and middle panels. The only difference between simulations is that in the right panel, two more rewards were earned before the red arrow. Therefore, the figure shows that although the probability of switching is higher after reward, it gets

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

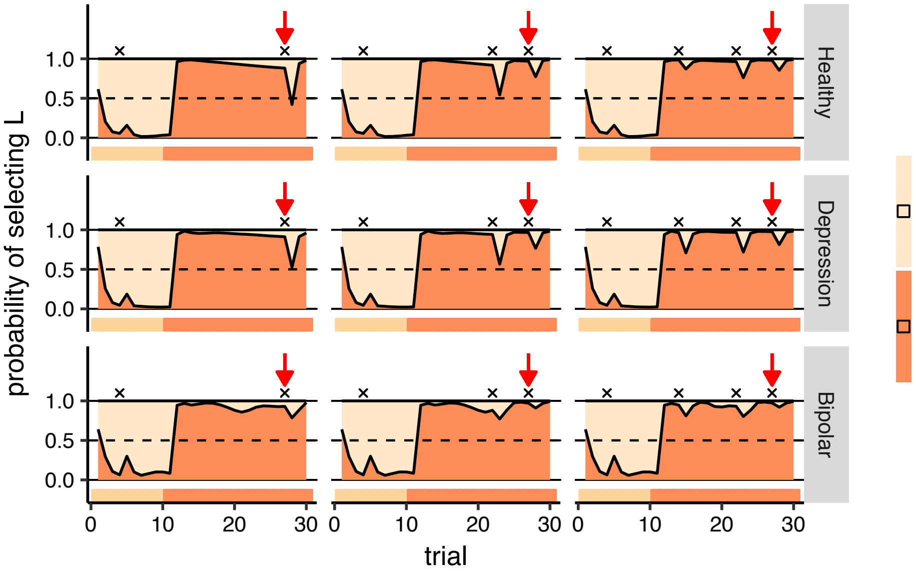

Figure 7. Off-policy simulations of RNN for all groups. Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph. Note that the simulation conditions is same as the one shown in Figure 6, and the first row here (HEALTHY group) is same as the first row shown in Figure 6 which is shown here again for the purpose of comparison with other groups.

smaller as more rewards are earned from an action. Indeed, this strategy makes subjects switch more from the inferior action as rewards are sparse on that action, and switch less from the superior action, as it is more frequently rewarded. This can reconcile the observations made in Figures 5, 4 that more responses were made on the better key, and at the same time, the probability of switching after reward was higher. As shown in Figure 6, GQL model produced a pattern similar to RNN, which is because this model tracks multiple values for each action (see Section S1 for details). Figure 7 shows the same simulations using RNN for all the groups (see Figures S5, S6, S7 for GQL, QLP and QL models respectively). Comparing the predictions at the red arrows for DEPRESSION and BIPOLAR groups, we see a pattern similar to HEALTHY group, although the differences are smaller in the BIPOLAR group (see Figure S9 for the effect of the initialisation of the model).

The above observations are consistent with the pattern of choices in empirical data as shown in Figure 8-left panel, which depicts the probability of staying on an action after earning reward as a function of how many rewards were earned after switching to the action (a similar graph using on-policy simulation of RNN is shown in Figure S11). In all the three groups, the probability of staying on an action (after earning a reward) was significantly higher when more than two rewards were earned previously (>2) compared to when no reward was earned (HEALTHY [$\beta = 0.148$, SE$= 0.037$, $p < 0.001$][9], DEPRESSION [$\beta = 0.188$, SE$= 0.045$, $p < 0.001$], BIPOLAR [$\beta = 0.150$, SE$= 0.056$, $p = 0.012$]), which is consistent with the behaviour of both RNN and GQL.

[9] The intercept was random-effect at the subject level; whether zero rewards or more than two rewards were earned previously was fixed-effect.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

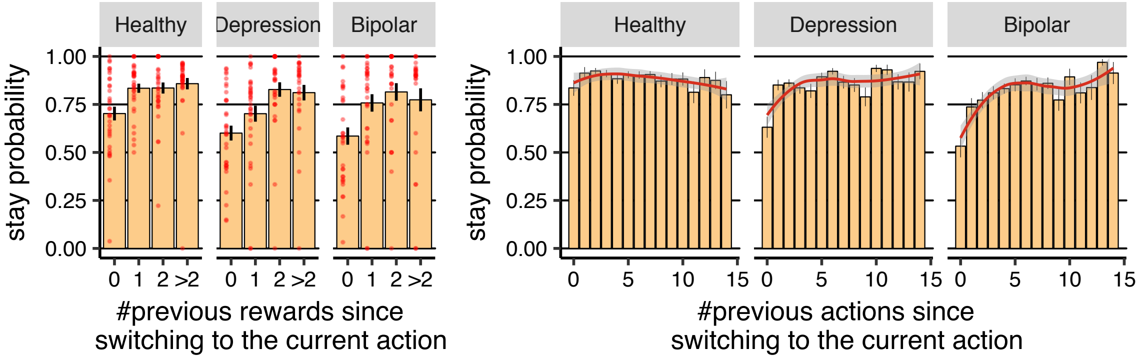

Figure 8. The effect of history of previous rewards and actions on future choices of the subjects. (Left-panel) Probability of staying on an action after earning reward as a function of number of actions taken since switching to the current action (averaged over subjects). Each red dot represents the data for each subject. **(Right-panel)** Probability of staying on an action as a function of number of actions taken since switching to the current action. The red line was obtained using Loess regression (Local Regression), which is a non-parametric regression approach. The grey area around the red line represents 95% confidence interval. Error-bars represent 1SEM.

The effect of repeating an action on choices. In the previous section we investigated the effect of previous rewards on choices. In this section we elaborate on how the history of previous actions affects current choices. Focusing on RNN simulations in the left-panel of Figure 6, an observation is that after switching to action L (after trial 10) the probability of staying on the action gradually decreases, i.e., although there is a high chance the next action will be similar to the previous one, subjects developed a tendency to make a switch the longer they stayed with an action. To compare this pattern with empirical data, we calculated the probability of staying on an action as a function of how many times the action was taken since switching, which is shown in Figure 8:right-panel[10](similar graphs for RNN on-policy simulations is shown in Figure S11). As the figure shows, for the HEALTHY group, the chance of staying on an action decreases as the action is taken more times [$\beta = -0.005$, SE= 0.001, $p = 0.006$][11], which is consistent with the behaviour of RNN. With regard to the baseline models, going back to Figure 6, we do not see a similar pattern, although in GQL there is a small decrement in the probability of staying on an action after earning the first reward.

Symmetric oscillations between actions. Next, we focus on RNN simulations for groups DEPRESSION and BIPOLAR in Figure 7. In the DEPRESSION group, the probability of staying on an action is almost flat with a slight decrement in the middle. For the BIPOLAR group, there is a dip around 10 trials after switching to action L (which will be around trial 20), and then the policy becomes flat. Referring to the empirical data, as shown in Figure 8:right-panel, the effect of number of actions in stay probabilities is not monotonic. In particular, as shown in Figure 8:right-panel, for DEPRESSION and BIPOLAR groups the probability of staying on an action immediately after switching to the action is around 50% - 60% (shown by the bar at $x = 0$ in Figure 8:right-panel), i.e., there is a 40% - 50% chance that the subject immediately switches back to the previous action. Based on this, we expect to see a 'dip'

[10] To be consistent with off-policy simulations, only trials in which (i) subjects did not earn a reward on that trial, (ii) subjects did not earn reward since switching to the current action, were included in the graph.

[11] The intercept was random-effect at the subject level; the number of times that an action was repeated since switching to the action was fixed-effect (between zero to 15 times). The dependent variable was the probability of staying on an action.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the simulations of DEPRESSION and BIPOLAR groups in Figure 7 just after switching to L action, which is not the case, pointing to an inconsistency between model predictions and empirical data.

To look closer at the above effect, we define a *run* of key presses as a sequence of presses on a certain key in a row, without switching to the other action[12]. Figure 9 shows the relationship between consecutive run lengths, i.e., the length of the current run of actions as a function of the length of the previous run of actions (see Figures S13, S14, and S15 for the similar graphs using on-policy simulations of RNN, GQL with $N = 2$ and GQL with $N = 10$, respectively). The dashed line in the figure indicates the points at which the current run length is the same the previous run length. Being close to this line implies that subjects are performing symmetric oscillations between the two actions, i.e., going back and forth between the two actions while performing an equal number of presses on each key. In particular, as the graph shows in the BIPOLAR group, and to an extent the in DEPRESSION group, a run of short length will trigger another run with a similar length. This implies that, if for example by chance a subject performs a run of length 1, it will initiate a sequence of oscillations between the two actions, which will keep the stay probabilities low during short runs, consistent with what we see at $x = 0$ in Figure 8:right-panel. This effect cannot be seen in the simulations that we showed in Figure 7, because the length of the previous run before switching to action L was 10 (there were 10 R actions), and therefore we do not expect the next run to be of length 1, neither do we expect to see a dip in policies just after the first switch.

As shown in Figure S10, majority of runs are of length 1 in the DEPRESSION, and BIPOLAR groups (around 17%, 37%, and 45% of runs are of length 1 in the HEALTHY, DEPRESSION, and BIPOLAR groups respectively). Given this, and the specific pattern of oscillations in the DEPRESSION and BIPOLAR groups, the next question is whether in the models a run of length 1 will trigger the oscillations, similar to the empirical data. We used a combination of off-policy and on-policy model simulations to answer this question. That is, during the off-policy phase we forced the model to make an oscillation between the two actions, and then after that we let the model select actions. We expect to see that in the HEALTHY group, the model will converge to one of the actions, but in DEPRESSION and BIPOLAR groups, we expect to see the initial oscillations trigger further switches. Simulations are presented in Figure 10, in which the sequence of actions fed to the model for the first 9 trials is (off-policy trials):

$$\mathrm{R, R, R, R, R, R, L, R, L,}$$

in which there are two oscillations at the tail of the sequence (R, L, R, L,). The rest of actions (trials 10-20) were selected based on which action the model assigns the highest probability[13]. As the simulation shows, at the beginning the probability that the model assigns to action R is high, but then after feeding the oscillations, the model predicts that the future actions will be oscillating in DEPRESSION and BIPOLAR groups, but not in HEALTHY group, consistent with what we expect to observe.

Therefore, RNN is able to produce symmetric oscillations and its behaviour is consistent with the subjects' actions. As Figure 10 shows, besides RNN, GQL was also able to produce length 1 oscillations to some extent (as shown for BIPOLAR group), which can be the reason that the prediction accuracy achieved by this model is significantly better than QLP in BIPOLAR and DEPRESSION groups (Figure 3) in

[12] For example, if the executed actions are L, R, R, L, then the length of the first of run is 1 (L), the length of second run is 2 (R, R), and the length of the third run is 1 (L).

[13] Note that in on-policy simulations, typically actions are selected probabilistically according to the probabilities that a model assigns to each action. However, in the on-policy simulations presented in this section in order to get consistent results across simulations, actions were *not* selected probabilistically, but they were chosen based on which actions get the highest prediction probability.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

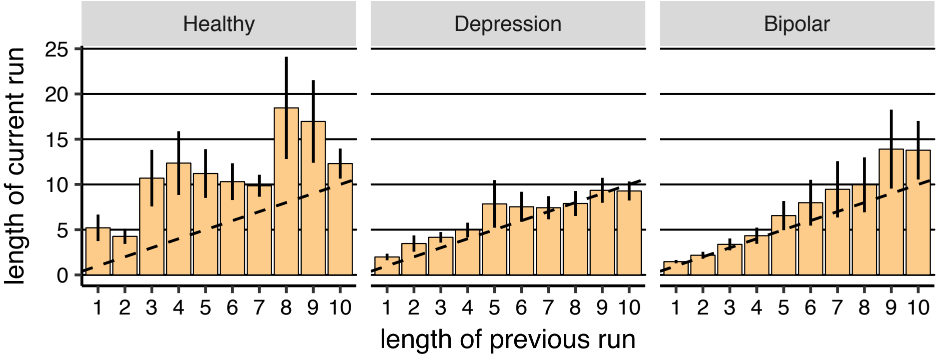

Figure 9. Median number of actions executed in a row before switching to another action (run of actions) in each subject as a function of length of previous run of actions (averaged over subjects). The dotted line shows the points in which the length of previous and current run are the same. Note that the use of median instead of average was because we aimed to illustrate most common 'length of current run', instead of average run length in each subject. Error-bars represent 1SEM.

which length 1 oscillations are more common (see Section S2 for more details). However, as shown in Figure S14, GQL failed to produce oscillations of longer lengths (even if we increase the capacity of GQL by using $N = 10$; see Figure S15), while RNN was able to do so (Figure S13). This inability in the GQL model is particularly problematic in DEPRESSION and HEALTHY groups, as these two groups tend to match the length of consecutive runs of actions. This could be the reason that in the cross-validation statistics RNN is significantly better than GQL in DEPRESSION and BIPOLAR groups.

Summary. Firstly, RNN was able to capture the immediate effect of rewards on actions (i.e., the 'dip' after rewards), as well as the effect of previous rewards on choices. GQL has the same ability, which enabled it to reproduce behavioural summary statistics shown in Figures 4, 5. Baseline reinforcement-learning models (QLP and QL) failed to capture either trend. Secondly, RNN was able to capture how choices change as an action is chosen repeatedly in a row, and also the symmetric oscillations between the actions, which GQL was unable to do so.

Discussion

Based on recurrent neural networks, we provide a method for learning a computational model that can characterize human learning processes in decision-making tasks. Unlike previous works, the current approach makes minimal assumptions about these learning processes; we showed that this agnosticism is important to be able to explain the data. In particular, subjects apparently used a mixture of different processes to select actions; there were some differences in these processes between healthy and the psychiatric groups. The RNN model was able to learn these processes from data. These processes were to a large extent inconsistent with Q-learning models, and were also rather hidden in the overall performance of the subjects in the task. This is an example of how our proposed framework can outperform previous approaches. Furthermore, we show that the model can be interpreted using off-policy simulations, providing insights into the learning processes used by humans. Finally, as an application of the model in computational psychiatry, we reported the performance of the model in predicting diagnostic labels.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

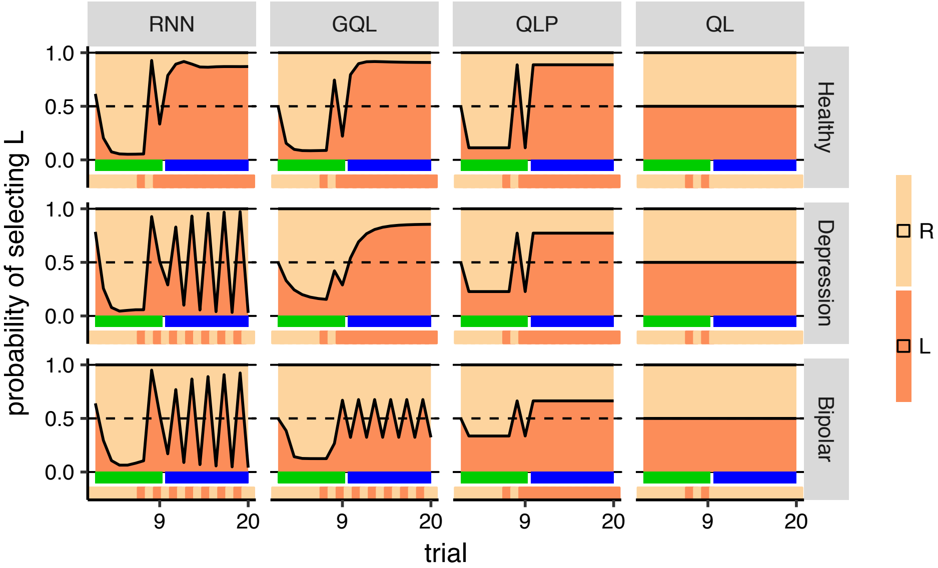

Figure 10. Mixed off-policy and on-policy simulations of the models. Each panel shows a simulation for 20 trials where the first nine trials is off-policy, and the next trials are on-policy during which the action with the highest probability is selected. Trials marked with the green ribbons are off-policy (actions are fed to the model), and the trials marked with the blue ribbons are on-policy (actions are selected by the model). The ribbon below each panel shows the action that was fed to the model (for the first 9 trials), and the action that was selected by the model (for the rest of trials). During off-policy trials, the sequence of actions that was fed to the model was R, R, R, R, R, R, L, R, L. See the text for the interpretation of the graph.

It might be possible to design different variants of Q-learning models (e.g., based on the analysis presented before) and obtain a more competitive prediction accuracy. For example, although it is non-trivial, one can design a new variant of GQL which is able to track the oscillatory behaviour. Our aim here was not to rule out this possibility, but to show that the strength of our analysis lies in its ability to automatically extract learning features – which were initially invisible in task performance metrics – from subjects' actions through learning to learn, without requiring feature engineering in the models. In this respect, although GQL was not used in the previous works, we designed and used it as the baseline model that can correctly capture high-level summary statistics of behaviour (Figures 5, 4), although misses deeper trends that are essential to characterize behaviour, particularly in psychiatric groups.

Indeed, our approach inherits this benefit from neural networks which have significantly simplified feature engineering in different domains (Lecun et al., 2015). However, our approach also inherits the black-box nature, i.e., the lack of interpretable working mechanism, of neural networks. This might not be an issue in some applications such as predicting diagnostic label of the subjects; however, it needs to be addressed in other applications in which the target of the study is obtaining an interpretable working mechanism. We did however show that running controlled experiments on the model through the off-policy simulations can provide some insights into the processes behind subjects' choices. Interpreting neural networks is an active area of research in machine learning (e.g., Karpathy et al., 2015), and the approach proposed here can benefit from further developments in this area. Even as a pure black-box model, the current approach can also contribute to the previous methods of computational modelling by providing a baseline for predictive accuracy. That is, as long as a candidate model does not provide better than or equal performance to RNN models, it means that there are certainly accessible behavioural

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trends that have been missed in the model structure. This is particularly important due to the natural randomness in human choices, making it unclear in many scenarios whether the model at hand (e.g., a Q-learning model) has reached the limit of predictability of choices, or whether it requires further improvements.

As we showed, off-policy simulations of the model can be used to gain insights into the model's working mechanism. However, off-policy simulations need to be designed manually to determine the inputs to the model. Here, we designed the initial off-policy simulations based on the specific questions and hypotheses that we were interested in and using overall behavioural statistics (Figure 6; Section S2). However, an important part of the behavioural processes, i.e., the tendency of subjects to oscillate between the actions, was not visible in those simulations, and because of this we designed another set of inputs to investigate the oscillations (Figure 10). This shows that the choices of off-policy simulations affect the interpretation of the model's working mechanism. As such, although RNN can be trained automatically and without intuition into the behavioural processes behind actions (e.g., Barak, 2017), the other part, i.e., designing off-policy simulations, is not automated and does need manual hypothesis generation. Automating this process requires a method that generates representative inputs (and network outputs) that most discriminately describe the differences between the psychiatric groups. We did not address this limitation in this work, and left it for future research.

Recurrent neural networks have been previously used to study reward-related decision-making (Song et al., 2017; Zhang et al., 2018), perceptual decision-making, performance in cognitive tasks, working-memory (Miconi, 2017; Carnevale et al., 2015; Mante et al., 2013; Song et al., 2016; Barak et al., 2013; Yang et al., 2017), motor patterns, motor reach and timing (Sussillo et al., 2015; Hennequin et al., 2014; Rajan et al., 2016; Laje and Buonomano, 2013). Typically, in these studies a RNN is trained based on the performance of the model in the task, which is different from the current study in which the aim of training is to generate a behaviour similar to the subjects', even if it leads to poor performance in the task. An exception is for example the study in Sussillo et al. (2015) in which a network was trained to generate outputs similar to electromyographic (EMG) signals recorded in behaving animals during a motor reach task. Interestingly, the study found that even though the model was trained purely based on EMG signals, the internal activity of the model resembled neural responses recorded from the motor cortex of the animals. A similar approach can be employed in future works to investigate whether brain activities during decision-making are related to the network activity.

With regard to predicting subjects' diagnostic labels, it is not surprising that the model was unable to achieve a high level of classification accuracy in predicting diagnostic labels. The reason is that there is a high level of heterogeneity in patients with the same diagnostic label, which for example is reflected in the wide variation in treatments and treatment outcomes in depression (e.g., Rush et al., 2006). Such variations might be reflected in the learning and choice abilities of the subjects, in which case, may be predicted using the model's inferred labels for each subject. On the other hand, purely as a diagnostic tool the current approach may help clinicians in situations that using questionnaires is not applicable (e.g., due to language/cultural barriers).

In the model fitting procedure used here, a single model was fitted to all subjects in each group, despite possible individual differences within a group. This was partly because we were interested in obtaining a single parameter set for making predictions for the held out subject in leave-one-out cross-validation experiments. That is, even if a mixed-effect model was fitted to the data, at the end, a summary of group statistics is required for making predictions about a new subject. In other applications one might be interested in estimating parameters for each individual (either network weights

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or parameters of the reinforcement-learning models); in this respect using a hierarchical model fitting procedure would be a more appropriate approach, which has been done previously for the reinforcement-learning models (e.g., Piray et al., 2014) and it would be an interesting future step to develop it for RNN models.

Along the same lines, a single RNN model, due to its rich set of parameters, might be able to learn and detect individual differences (e.g., differences in the learning-rates of subjects) at early trials of the task, and then use this information for making predictions for later trials. For example, in the learning-to-learn phase, the model might learn that subjects either have a very high, or a very low learning-rate. Then, when being evaluated in the actual learning task, the model can use its observations from subjects' choices at early trials to determine whether the learning-rate for that specific subject is high or low, and then utilise that information for making more accurate predictions in latter trials. Determining such individual-specific traits in early trials of the task is presumably *not part of the computational processes* occurring in the subject's brain during the task, but it is occurring in the model merely to make more accurate predictions. Therefore, if the network learns to do so, it might not be straightforward to treat such models as the computational models for subject's choices, but only as the models that are able to make predictions for the choices.

Acknowledgments

PD was partly funded by the Gatsby Charitable Foundation. PD is currently on a leave of absence at Uber Technologies. Neither body influenced this study. AD was supported by by grant DP150104878 from the Australian Research Council. BWB was supported by a Senior Principal Research Fellowship from the National Health & Medical Research Council of Australia, GNT1079561.

References

Jerome R Busemeyer and Julie C Stout. A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. *Psychological assessment*, 14(3): 253–62, 2002.

Amir Dezfouli, Mohammad Mahdi Keramati, Hamed Ekhtiari, H. Safaei, and Caro Lucas. Understanding Addictive Behavior on the Iowa Gambling Task Using Reinforcement Learning Framework. In *30th Annual Conference of the Cognitive Science Society*, pages 1094–1099, 2007.

P Read Montague, Raymond J Dolan, Karl J Friston, and Peter Dayan. Computational psychiatry. *Trends in cognitive sciences*, 16(1):72–80, jan 2012.

Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–9, jun 2006.

C.J.C.H. Watkins. *Learning from Delayed Rewards*. Ph.d. thesis, Cambridge University, 1989.

Payam Piray, Yashar Zeighami, Fariba Bahrami, Abeer M Eissa, Doaa H Hewedi, and Ahmed A Moustafa. Impulse control disorders in Parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation. *The Journal of neuroscience*, 34(23):7814–24, 2014.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv preprint arXiv:1611.02779*, 2016.

Ari Weinstein and Matthew M Botvinick. Structure Learning in Motor Control: A Deep Reinforcement Learning Model. *arXiv preprint arXiv:1706.06827*, 2017.

Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1): 56, 1960.

R C Young, J T Biggs, V E Ziegler, and D A Meyer. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry*, 133(5):429–435, 1978.

Howard H Goldman, Andrew E Skodol, and Tamara R Lave. Revising axis V for DSM-IV: a review of measures of social functioning. *Am J Psychiatry*, 149(9):1148–1156, 1992.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and Others. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Brian Lau and Paul W Glimcher. Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, 84(3):555–79, 2005.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models*, 2016.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 14764687.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, 46:1–6, 2017. ISSN 18736882. doi: 10.1016/j.conb.2017.06.003.

H. Francis Song, Guangyu R. Yang, and Xiao Jing Wang. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 6:1–24, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Zhewei Zhang, Zhenbo Cheng, Zhongqiao Lin, Chechang Nie, and Tianming Yang. A neural network model for the orbitofrontal cortex and task space acquisition during reinforcement learning. *PLOS Computational Biology*, 14(1):e1005925, 2018. ISSN 1553-7358.

Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, 6:1–24, 2017.

Federico Carnevale, Victor DeLafuente, Ranulfo Romo, Omri Barak, and Néstor Parga. Dynamic Control of Response Criterion in Premotor Cortex during Perceptual Detection under Temporal Uncertainty. *Neuron*, 86(4):1067–1077, 2015.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

H. Francis Song, Guangyu R. Yang, and Xiao Jing Wang. Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework. *PLoS Computational Biology*, 12(2):1–30, 2016.

Omri Barak, David Sussillo, Ranulfo Romo, Misha Tsodyks, and L. F. Abbott. From fixed points to chaos: Three models of delayed discrimination. *Progress in Neurobiology*, 103:214–222, 2013.

Guangyu Robert Yang, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Clustering and compositionality of task representations in a neural network trained to perform many cognitive tasks. *bioRxiv*, page 183632, 2017.

David Sussillo, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18(7): 1025–1033, 2015.

Guillaume Hennequin, Tim P. Vogels, and Wulfram Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406, 2014.

Kanaka Rajan, Christopher D D. Harvey, and David W W. Tank. Recurrent Network Models of Sequence Generation and Memory. *Neuron*, 90(1):128–142, 2016.

Rodrigo Laje and Dean V. Buonomano. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, 16(7):925–933, 2013.

A. John Rush, Madhukar H. Trivedi, Stephen R. Wisniewski, Andrew A. Nierenberg, Jonathan W. Stewart, Diane Warden, George Niederehe, Michael E. Thase, Philip W. Lavori, Barry D. Lebowitz, Patrick J. McGrath, Jerrold F. Rosenbaum, Harold A. Sackeim, David J. Kupfer, James Luther, and Maurizio Fava. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry*, 163(11):1905–1917, 2006.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supporting information

S1 Behavioural analysis using GQL

In GQL simulations presented in Figure 6, an observation is that earning a reward (shown by black crosses) causes a 'dip' in the probability of staying on an action, which shows the tendency to switch to the other action. This is consistent with the observation made in Figure 5 that the probability of switching increases after rewards. The reason that GQL is able to produce different predictions to that of QL and QLP is that in this model, the contribution of action values to choices can be negative, i.e., higher values can lead to a lower probability of staying on an action. Indeed, examining the learned parameters for this model (Table S4) revealed that for each action, values are updated at two different rates, a slow rate (0.145) and a fast rate (0.815), and the coefficient for the values that are updated at the faster rate is negative (-1.002). This implies that after earning a reward the value of the action taken increases quickly, but that increase leads to the lower probability of selecting the action - which makes the 'dip' in policies following the rewards.

The next observation was with respect to the effect of previous rewards on the probability of switching after a reward. As shown in Figure 6, GQL model produced a pattern similar to RNN, and it can be seen that the probability of staying on an action increases after earning a reward, causing the depth of the dip after earning reward to become smaller as more rewards are earned. This ability of GQL is because this model tracks two values for each action, one of them updated with a fast learning-rate and the other with a slow learning-rate. The one that updates faster plays a role in the dip following each reward. On the other hand, the value that updates slower (at learning-rate 0.145) has an opposite effect since it contributes to choices with a positive coefficient (4.258), and therefore, with increases in the value after reward the probability of staying with an action increases. Based on this, allowing the model to track two different values for each action is important, and the model will not be able to produce this behaviour if it tracks only one value for each action ($N = 1$) as shown in Figure S8.

S2 The choice of off-policy settings

In the simulations shown in Figure 6, action R is fed into the model for the first 10 trials before switching to the other action. This is based on the fact that in the empirical data, the average length of staying with an action (when one reward is earned in the middle of the execution of the action) is 9.8. The first, second and third rewards in Figure 6 are delivered after an action was taken 4, 12, and 17 times respectively. This is based on the fact that in the empirical data, the average number of key-presses in order to earn the first, second and third rewards is 4.07, 11.6, and 17.4 respectively.

In the simulations shown in Figure 10, the reason for adding leading R before oscillations is to show that the models do not oscillate all the time, but only after they are fed with oscillations. Indeed, QLP is in principle able to produce 1-step oscillations (singe-action runs) by assigning a negative weight to the perseveration parameter, i.e., instead of the model having a tendency to stay on the previously selected action, it will have a tendency not to stay on the selected action. However, under this condition the model will keep oscillating between the actions from trial 1, implying that it can only produce runs of length 1 no matter what the length of the previous run of actions was, which is inconsistent with the empirical data presented in Figure 9.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

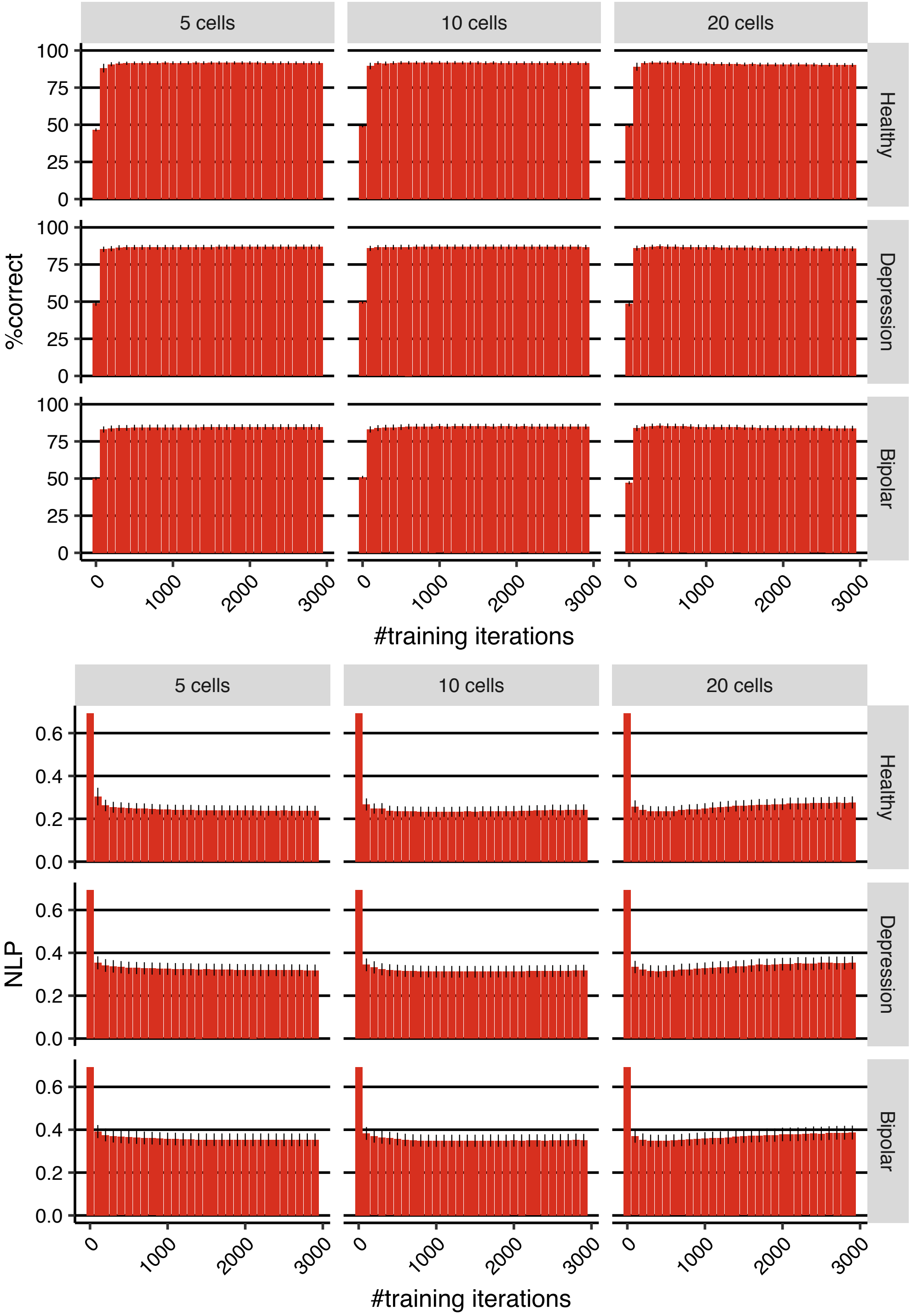

Figure S1. Cross-validation results for different number of cells and optimization iterations. **(Top-panel)** Percentage of actions predicted correctly averaged over leave-one-out cross-validation folds. **(Bottom-panel)** Mean NLP averaged over cross-validation folds. Error-bars represent 1SEM.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

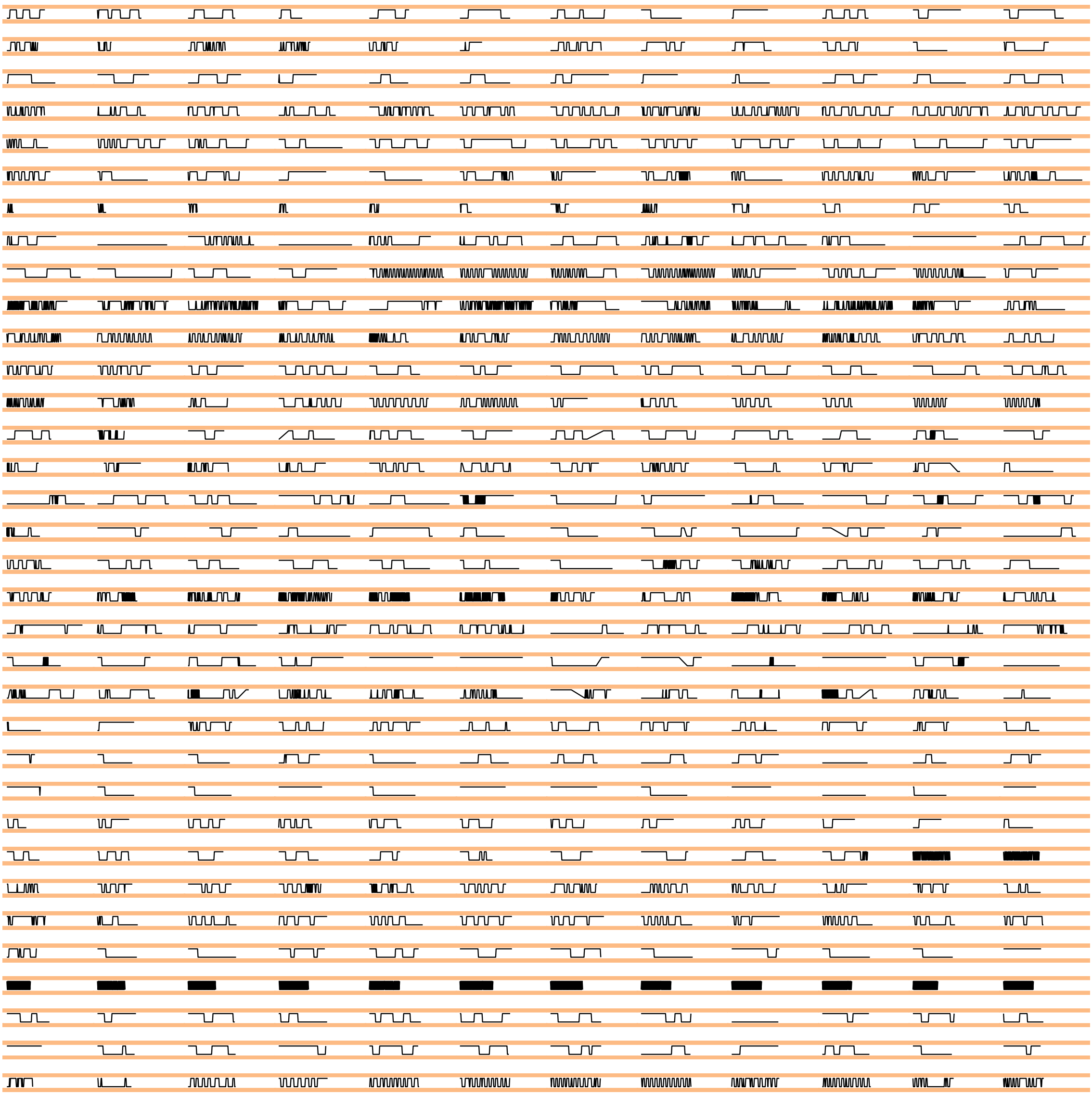

Figure S2. Choices in HEALTHY group. Each row shows choices of a subject across different blocks (12 blocks).

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

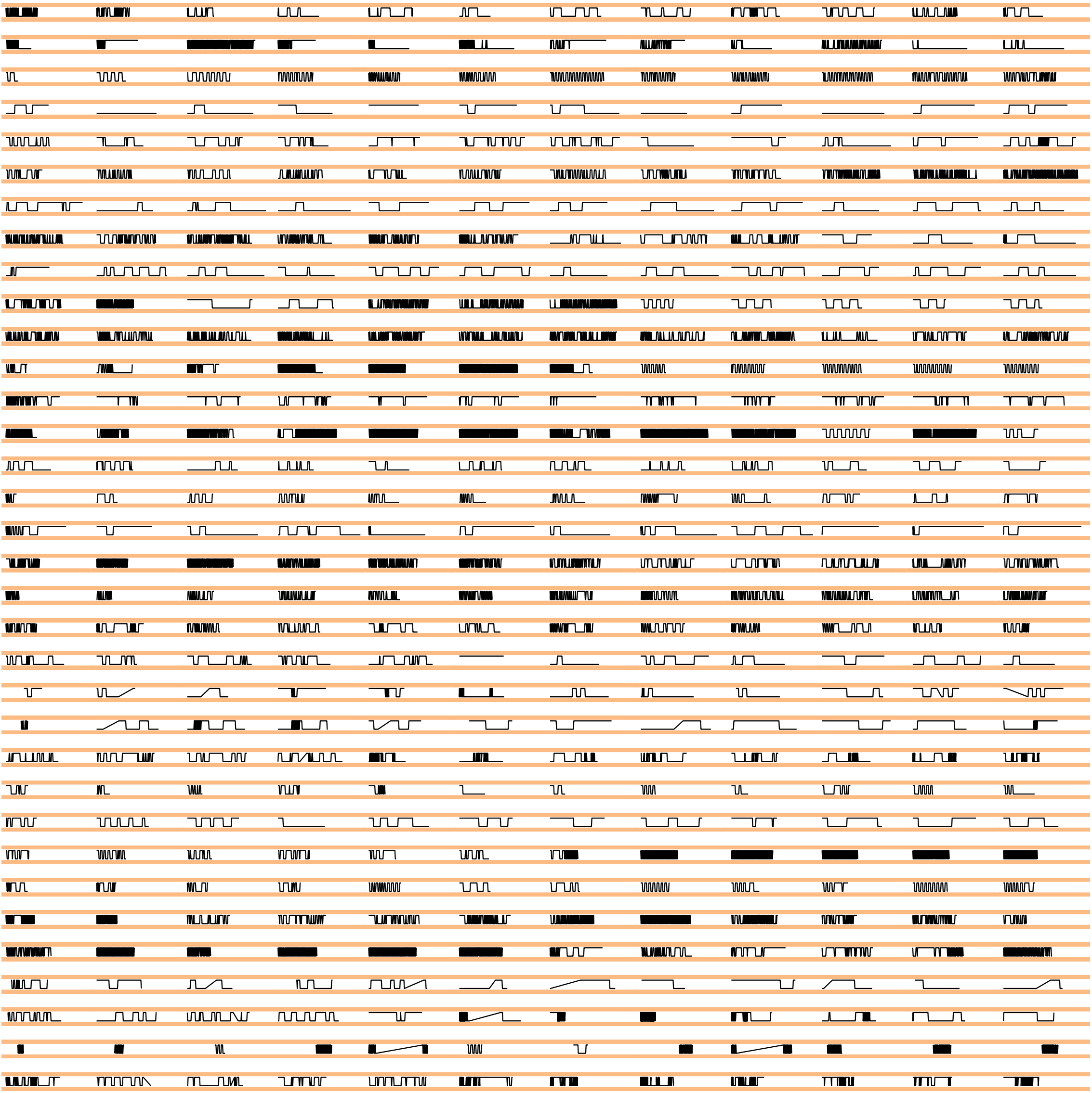

Figure S3. Choices in DEPRESSION group. Each row shows choices of a subject across different blocks (12 blocks).

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

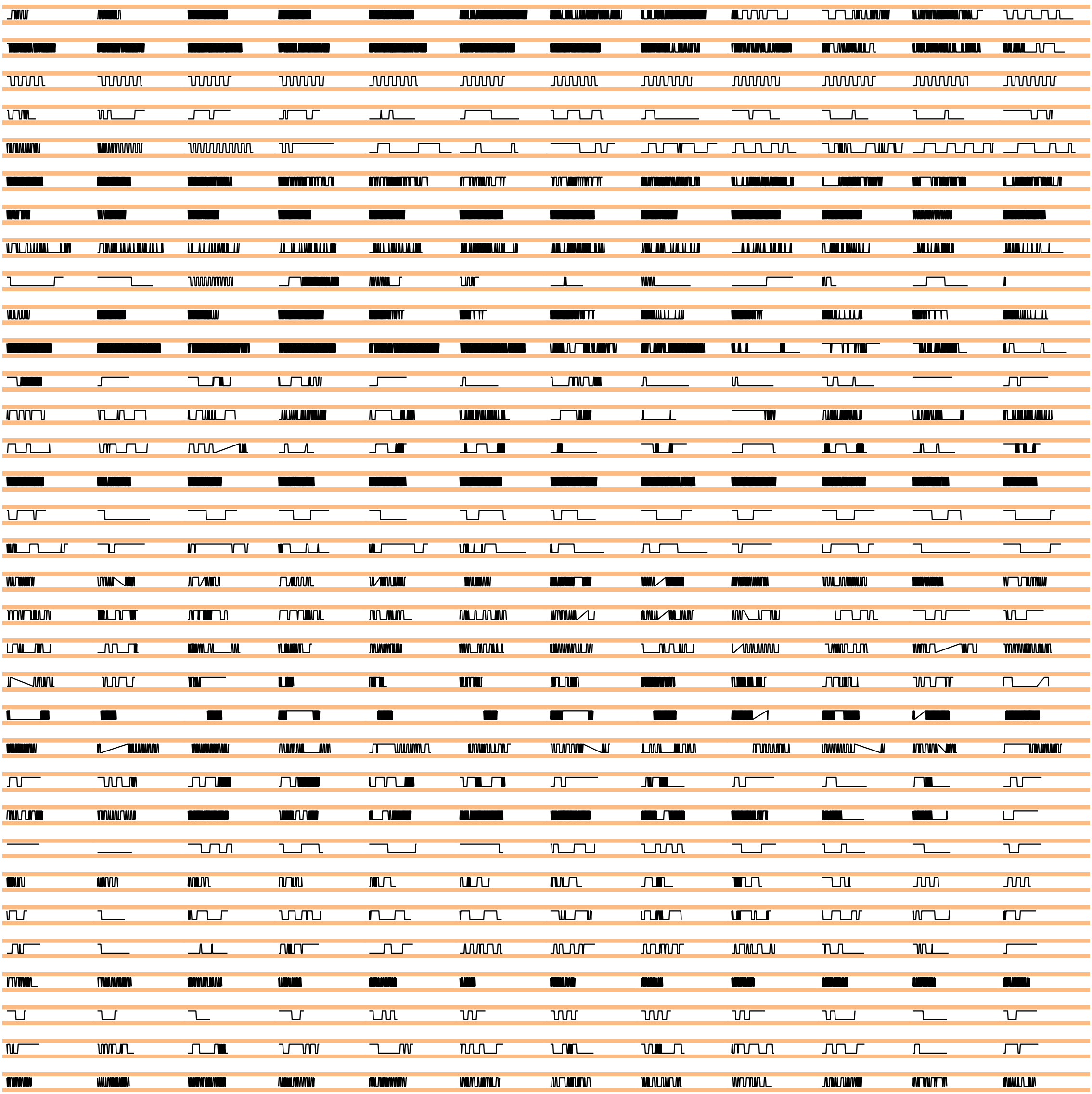

Figure S4. Choices in BIPOLAR group. Each row shows choices of a subject across different blocks (12 blocks).

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

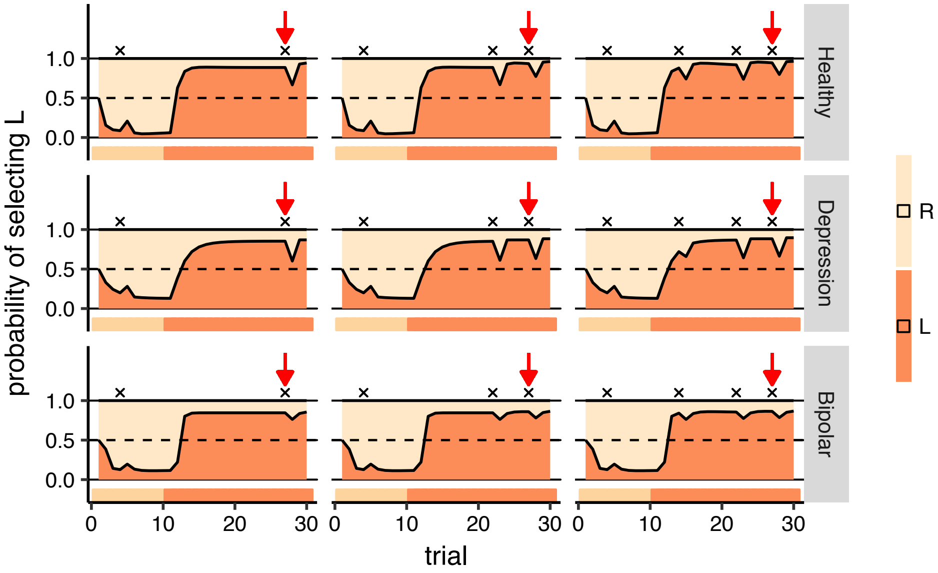

Figure S5. Off-policy simulations of GQL ($N = 2$). Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph. Note that the simulation conditions is same as the one depicted in Figures 7 and 6

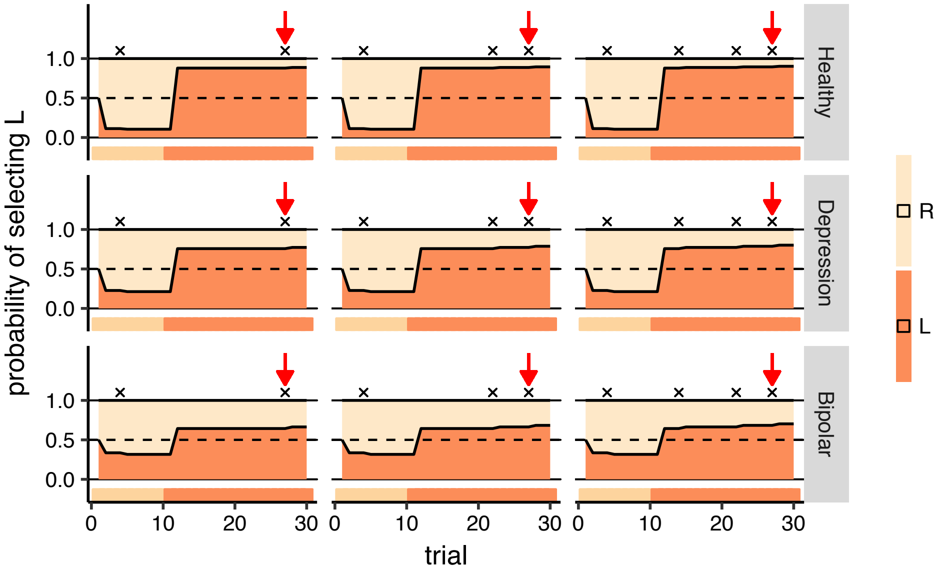

Figure S6. Off-policy simulations of QLP. Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph. Note that the simulation conditions is same as the one depicted in Figures 7 and 6

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

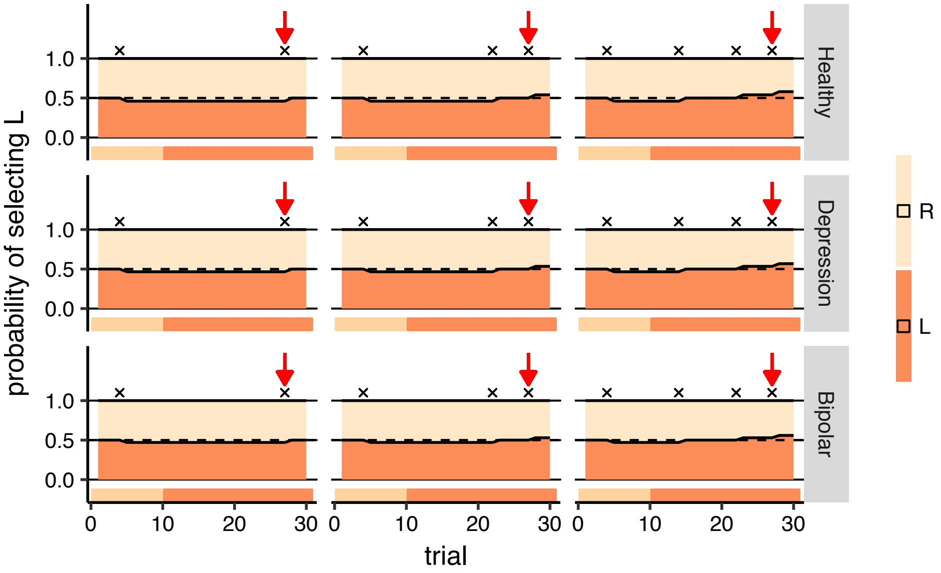

Figure S7. Off-policy simulations of QL. Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph. Note that the simulation conditions is same as the one depicted in Figures 7 and 6

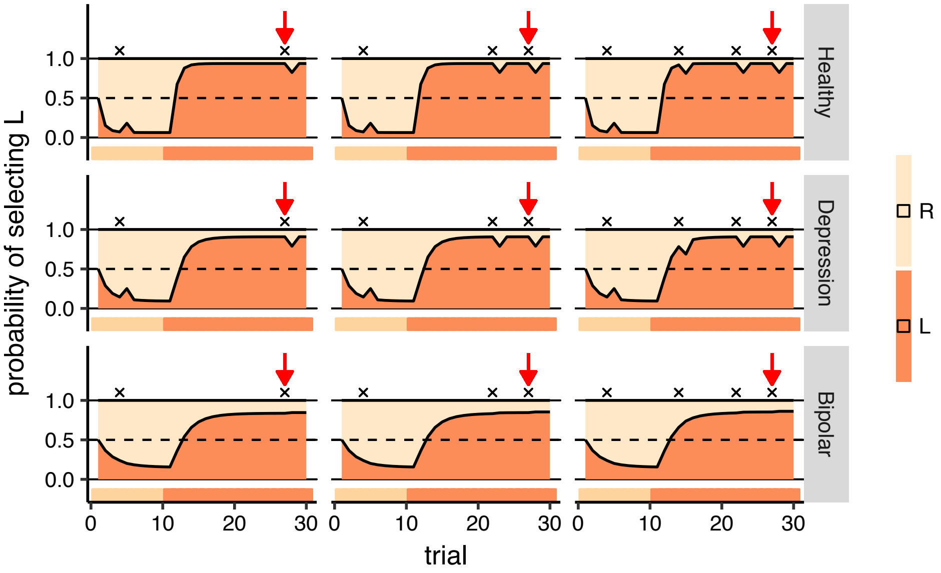

Figure S8. Off-policy simulations of GQL with $N = 1$. Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph. Note that the simulation conditions is same as the one depicted in Figures 7 and 6

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

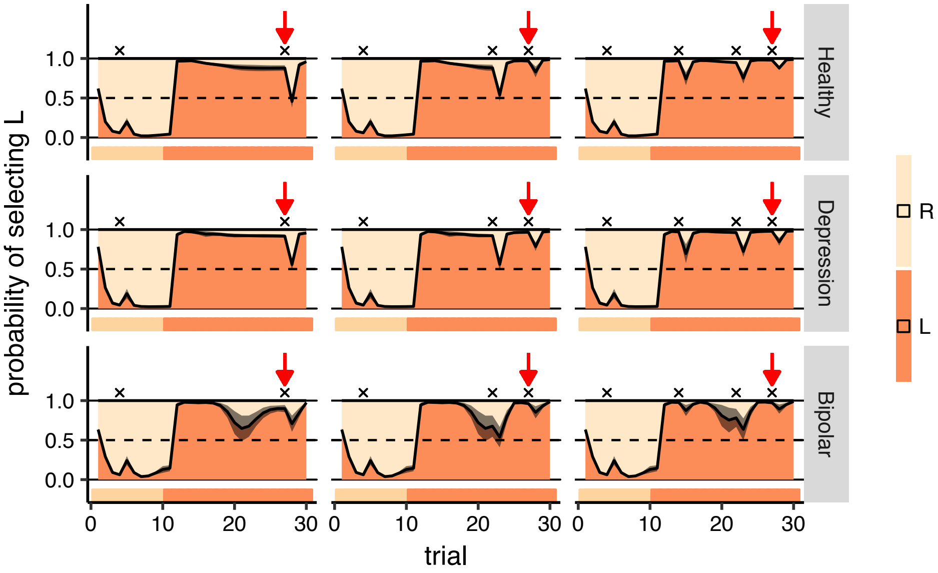

Figure S9. The effect of the initialisation of the network on the off-policy simulations of RNN. The simulation conditions are the same as the ones depicted in Figures 7 and 6. Here, 15 different initial networks were generated and optimised and the policies of the models at each trial were averaged. The gray ribbon around the policy shows the standard deviation of the policies. Each panel shows a simulation for 30 trials (horizontal axis), and the vertical axis shows the predictions of each model on each trial. The ribbon below each panel shows the action which was fed to the model on each trial. In the first 10 trials, the action that the model received was R and in the next 20 trials it was L. Rewards are shown by black crosses (x) on the graphs. See text for the interpretation of the graph.

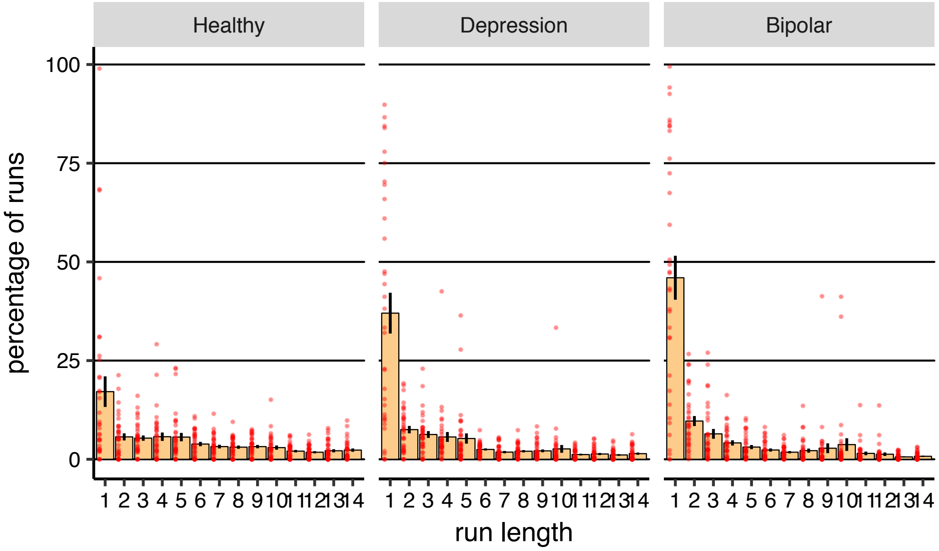

Figure S10. Percentage of each run of actions relative to the total number of runs for each subject. Percentage of each length of run of actions relative to the total number of run of actions in each subject (averaged over subjects). Red dots represent data for each subject, and error-bars represent 1SEM.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

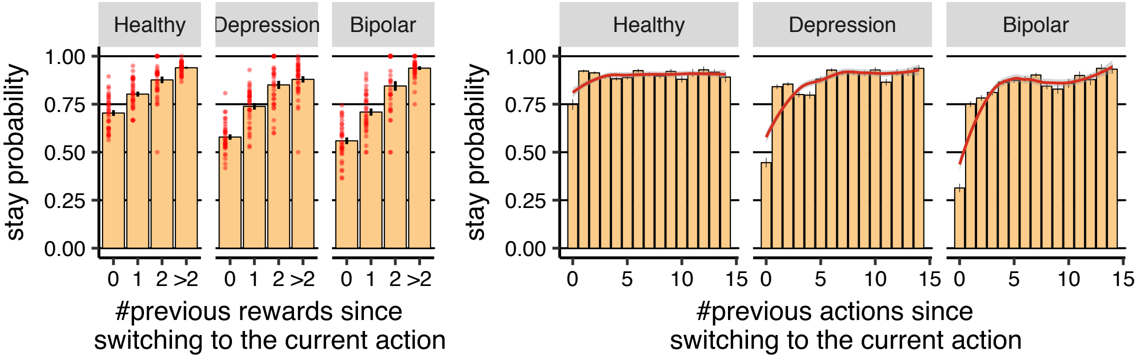

Figure S11. **RNN simulations**. The graph is similar to Figure 8 but using data from RNN simulations. **(Left-panel)** Probability of staying on an action after earning reward as a function of number of actions taken since switching to the current action (averaged over subjects). Each red dot represents the data for each subject. **(Right-panel)** Probability of staying on an actions as a function of number of actions taken since switching to the current action. The red line was obtained using Loess regression (Local Regression), which is a non-parametric regression approach. The grey area around the red line represents 95% confidence interval. Error-bars represent 1SEM.

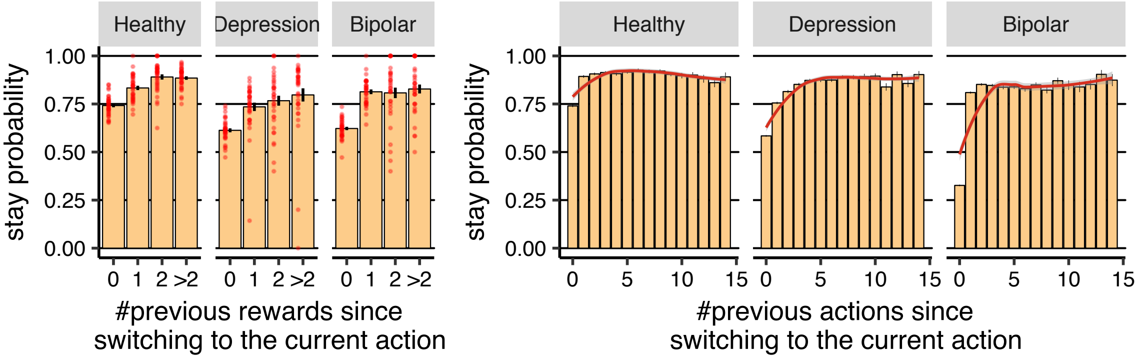

Figure S12. **GQL simulations** ($N = 2$). The graph is similar to Figure 8 but using data from GQL simulations with $N = 2$. **(Left-panel)** Probability of staying on an action after earning reward as a function of number of actions taken since switching to the current action (averaged over subjects). Each red dot represents the data for each subject. **(Right-panel)** Probability of staying on an actions as a function of number of actions taken since switching to the current action. The red line was obtained using Loess regression (Local Regression), which is a non-parametric regression approach. The grey area around the red line represents 95% confidence interval. Error-bars represent 1SEM.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

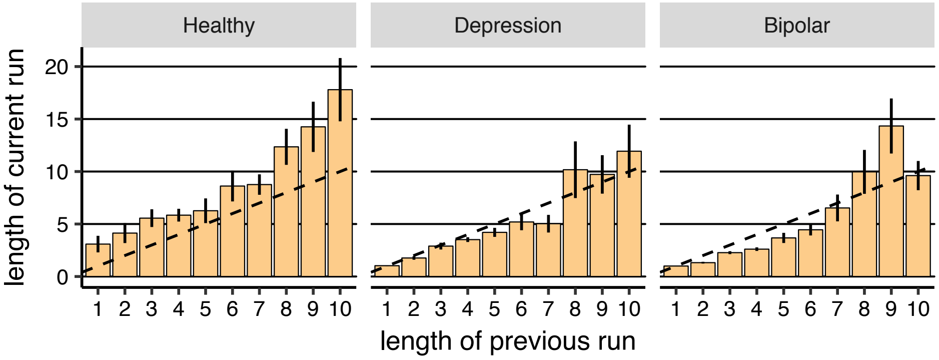

Figure S13. RNN simulations. The graph is similar to Figure 9 but using data from RNN simulations. Median number of actions executed in a row before switching to another action (run of actions) in each subject as a function of length of previous run of actions (averaged over subjects). The dotted line shows the points in which the length of previous and current run are the same. Note that the use of median instead of average was because we aimed to illustrate most common 'length of current run', instead of average run length in each subject. Error-bars represent 1SEM.

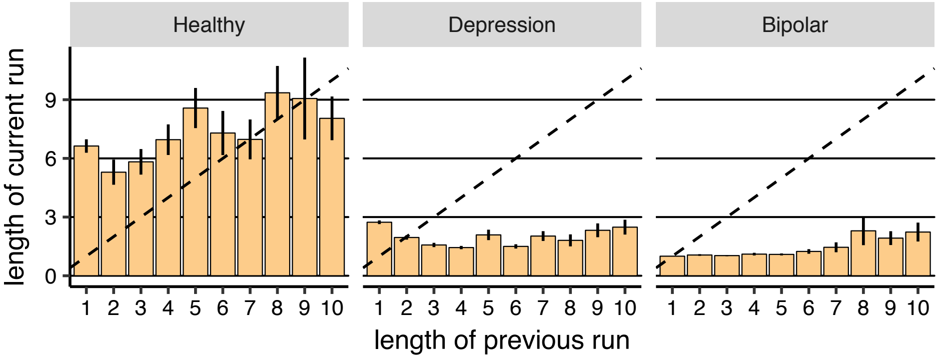

Figure S14. GQL simulations ($N = 2$). The graph is similar to Figure 9 but using data from GQL simulations with $N = 2$. Median number of actions executed in a row before switching to another action (run of actions) in each subject as a function of length of previous run of actions (averaged over subjects). The dotted line shows the points in which the length of previous and current run are the same. Note that the use of median instead of average was because we aimed to illustrate most common 'length of current run', instead of average run length in each subject. Error-bars represent 1SEM.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

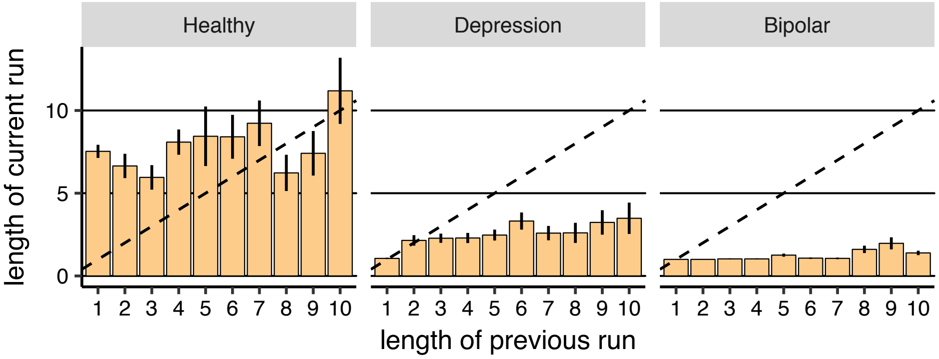

Figure S15. **GQL simulations ($N = 10$).** The graph is similar to Figure 9 but using data from GQL simulations with $N = 10$. Median number of actions executed in a row before switching to another action (run of actions) in each subject as a function of length of previous run of actions (averaged over subjects). The dotted line shows the points in which the length of previous and current run are the same. Note that the use of median instead of average was because we aimed to illustrate most common ‘length of current run’, instead of average run length in each subject. Error-bars represent 1SEM.

bioRxiv preprint doi: https://doi.org/10.1101/285221; this version posted March 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S1. Prediction of diagnostic labels using GQL ($N = 2$). Number of subjects for each true- and predicted-labels. The numbers inside parenthesis are the percentage of number subjects relative to the total number of subjects in each diagnostic group.

		predicted labels		
		HEALTHY	DEPRESSION	BIPOLAR
true labels	HEALTHY	29 (85%)	2 (5%)	3 (8%)
	DEPRESSION	16 (47%)	7 (20%)	11 (32%)
	BIPOLAR	12 (36%)	6 (18%)	15 (45%)

Table S2. Estimated parameters for QL model.

	α	β	$\alpha\beta$
HEALTHY	0.0002	616.559	0.162
DEPRESSION	0.00001	11760.3	0.14
BIPOLAR	0.00002	4025.94	0.12

Table S3. Estimated parameters for QLP model.

	α	β	κ	$\alpha\beta$
HEALTHY	0.0008	98.8895	2.0634	0.079
DEPRESSION	0.00003	2479.31	1.223	0.09
BIPOLAR	0.00008	1113.97	0.680	0.09

Table S4. Estimated parameters for GQL model with $N = 2$.

	$\mathbf{\Phi}$	$\mathbf{\Psi}$	$\mathbf{B}$	$\mathbf{K}$	$\mathbf{C}$
HEALTHY	[0.145 0.815]	[0.635 0.389]	[4.258 -1.002]	[3.268 -0.974]	[[-14.256 4.243] [17.998 -6.335]]
DEPRESSION	[0.003 0.999]	[0.399 0.3199]	[8.691 -0.315]	[1.709 0.077]	[[14.918 6.112] [19.599 -7.292]]
BIPOLAR	[0.147 0.654]	[0.897 0.999]	[4.363 -1.1453]	[14.447 -12.501]	[[0.174 -15.199] [-2.936 15.164]]

Table S5. Negative log-likelihood for each model optimized over all the subjects in each group.

	RNN	GQL	QLP	QL
HEALTHY	9421.6660	12939.40	14557.44	27616.79
DEPRESSION	13158.1074	19735.61	23378.65	29862.19
BIPOLAR	12891.3496	19363.08	24859.15	26843.88

Table S6. Negative log-likelihood for each model. For RNN a single model was fitted to the whole group using ML estimation. For baseline methods (GQL, QLP, and QL), a separate model was fitted to each subject, and the reported number is the some of negative log-likelihoods over the whole group.

	RNN	GQL	QLP	QL
HEALTHY	9421.6660	9482.846	11529.58	26080.13
DEPRESSION	13158.1074	14668.763	17837.02	28448.94
BIPOLAR	12891.3496	14157.206	16912.37	25874.56

Table S7. Mean and standard deviation of negative log-likelihood for RNN over 15 different initialisations of the model and optimised over all the subjects in each group.

	mean (standard deviation)
HEALTHY	9450.937 (81.512)
DEPRESSION	13268.186 (135.499)
BIPOLAR	12885.629 (134.552)