Blaming blunders on the brain: can indifferent choices be driven by range adaptation or synaptic plasticity?

Jules Brochard[1,2,3], Jean Daunizeau[1,2,3]

[1] Sorbonne Université, Paris, France

[2] Institut du Cerveau, Paris, France

[3] INSERM UMR S1127

Address for correspondence:

Jean Daunizeau

Motivation, Brain, and Behavior Group

Paris Brain Institute

47, boulevard de l'Hôpital, 75013, Paris, France

Tel: +33 1 57 27 47 19

E-mail: jean.daunizeau@gmail.com

Keywords: fMRI, representational similarity analysis, artificial neural networks, decision, risk.

Abstract

Computational investigations of learning and decision making suggest that systematic deviations to adaptive behavior may be the incidental outcome of biological constraints imposed on neural information processing. In particular, recent studies indicate that range adaptation, i.e., the mechanism by which neurons dynamically tune their output firing properties to match the changing statistics of their inputs, may drive plastic changes in the brain's decision system that induce systematic deviations to rationality. Here, we ask whether behaviorally-relevant neural information processing may be distorted by other incidental, hard-wired, biological constraints, in particular: Hebbian plasticity. One of our main contributions is to propose a simple computational method for identifying (and comparing) the neural signature of such biological mechanisms or constraints. Using ANNs (i.e., artificial neural network models) and RSA (i.e., representational similarity analysis), we compare the neural signatures of two types of hard-wired biological mechanisms/constraints: namely, range adaptation and Hebbian plasticity. We apply the approach to two different open fMRI datasets acquired when people make decisions under risk. In both cases, we show that although peoples' apparent indifferent choices are well explained by biologically-constrained ANNs, choice data alone does not discriminate between range adaptation and Hebbian plasticity. However, RSA shows that neural activity patterns in bilateral Striatum and Amygdala are more compatible with Hebbian plasticity. Finally, the strength of evidence for Hebbian plasticity in these structures predicts inter-individual differences in choice inconsistency.

Introduction

Why do we overreact to emotional stimuli? Why are our judgments plagued with errors and biases? Why do we engage in behaviors whose consequences may be detrimental? That the brain's biology is to blame for all kinds of cognitive and/or behavioural flaws is not a novel idea (Buschman et al., 2011; Marois and Ivanoff, 2005; Miller and Buschman, 2015; Ramsey et al., 2004). However, providing neuroscientific evidence that a hard-wired biological constraint shapes and/or distorts the way the brain processes information is not an easy task. This is because whether the brain deviates from how it *should* process a piece of information is virtually unknown. In this work, we show how one may use multivariate analysis of fMRI data to identify the neural signature of incidental, hard-wired, biological constraints on behaviorally-relevant neural information processing.

Over the past two decades, cognitive neuroscience has involved much effort into developing computational means to understand how the brain processes information. In particular, the computational neuroscience of perception, learning, and decision making has now reached a stage of maturity, both in terms of its methods and models and in terms of the reproducibility of the ensuing results. For example, neuroscientific evidence that basal ganglia encode the reward prediction error that enables reinforcement learning (i.e., learning from reward feedbacks) has been found repetitively in monkeys (Fiorillo et al., 2003; Schultz et al., 1997) and humans (Abler et al., 2006; Diederen et al., 2016; Garrison et al., 2013). From a methodological standpoint, this line of study is remarkable for two reasons. First, it highlights the importance of behavioral measurements for understanding how the brain processes information. This shifts the scientific question from identifying how the brain *encodes* incoming information (e.g., cues and feedbacks) to assessing how it *uses* this information to produce behavioral responses. Second, its theoretical basis is derived from formal computational models of learning originating from research in the field of artificial intelligence and robotics (Dayan and Daw, 2008; Sutton and Barto, 1998). This provides a formal reference point for interpreting neural signals in

terms of neural *computations*, i.e., intermediary steps in neural information processing geared towards producing adapted behavioral responses.

Taken in isolation, none of these two aspects is particularly novel. Retrospectively, the focus on brain-behavior relationships is the hallmark of behavioral neuroscience. And computational neuroscience already had enabled deep quantitative insights for understanding the neural code of perceptual and motor systems, providing unprecedented empirical evidence for, e.g., population coding (Averbeck et al., 2006; Georgopoulos et al., 1986), predictive coding (Bastos et al., 2012; Hosoya et al., 2005) or efficient coding (Barlow, 1961; Lewicki, 2002). But in combination, these two aspects allow one to understand how brain computations eventually shape non-trivial behavior. This has typically be done in two different ways. On the one hand, one may look for neural evidence of cognitive mechanisms that provide candidate explanations for observed behavioral deviations to normative theories. For example, this approach has placed the putative distortions of prospective loss perceptions that drive irrational risk attitudes on a firm neuroscientific footing (Martino et al., 2006, 2010; Tom et al., 2007). Critically, this line of work typically also demonstrates the relevance of neural data for understanding inter-individual differences w.r.t. the magnitude of behavioral distortions. For example, it was shown that those people who exhibit a strong optimism bias are those people whose encoding of disappointing prediction errors (in the right frontal gyrus) was the weakest (Sharot, 2011; Sharot et al., 2011). On the other hand, one may disclose non-trivial behavioral consequences of the computational properties of neural information processing. For example, it was shown that the brain's reliance on efficient coding induced systematic biases in both perceptual and value-based decisions (Louie and Glimcher, 2012; Polanía et al., 2019; Soltani et al., 2012; Wei and Stocker, 2015; Zimmermann et al., 2018). The irony here is that efficient coding is the brain's optimal solution to the problem of building reliable cognitive representations under limited neural resources (Barlow, 1961; Simoncelli and Olshausen, 2001). In brief, this series of work provides evidence for the impact of biological constraints on behaviorally-relevant information processing.

One critical insight here was that efficient coding induces plastic changes in the brain's decision system that was incidental, i.e., they were not instrumental to the decision task (Conen and Padoa-Schioppa, 2019). More precisely, the encoding of value in OFC neurons was shown to obey a ubiquitous, hard-wired, biological constraint, namely: range adaptation (Burke et al., 2016; Cox and Kable, 2014; Elliott et al., 2008; Kobayashi et al., 2010; Padoa-Schioppa, 2009). Range adaptation is the mechanism by which neurons dynamically tune their output firing properties to match the changing statistics of their inputs, hence implementing efficient coding under the constraint of bounded neural activation range (Brenner et al., 2000; Laughlin, 1981; Wark et al., 2007). Although a major breakthrough in decision neuroscience, these studies suffer from two methodological weaknesses. First, they rely on a normative reference model that describes how the brain should process behaviorally-relevant information, whose computational properties are altered by range adaptation. In turn, neuroscientific evidence for range adaptation is mostly indirect because it relies on validating its corollary consequence in terms of value distortions (e.g., divisive normalization), rather than identifying its neural signature (but see Zimmermann et al., 2018). Second, other alternative computational mechanisms that may make qualitatively similar predictions are ignored. In particular, one may argue that many forms of plasticity may, in principle, induce dynamic changes in the brain's decision circuits that may eventually be confounded with range adaptation. A ubiquitous and ever-persistent example of this is Hebbian synaptic plasticity (Hebb, 1950), which is central to, e.g., development and recovery from injury (Fox and Stryker, 2017; Martens et al., 2015; Turrigiano, 2017). A plethora of electrophysiological studies have established its many variants, including, but not limited to, spike-timing dependent plasticity and long-term potentiation/depression (Fox and Stryker, 2017; Lisman, 2017; Shouval et al., 2010; Zenke and Gerstner, 2017). Critically, Hebbian plasticity does not reduce to range adaptation, and one may reasonably ask which of these two hard-wired mechanisms is the most constraining for behaviorally-relevant neural information processing.

This work is a first step towards solving the two above issues. In brief, we propose a computational method for identifying (and comparing) the neural signature of biological mechanisms or constraints

on behaviorally-relevant neural information processing. We bypass the issue of defining a normative reference model for neural information processing by fitting ANNs (i.e., artificial neural network models) to behavioral data, with and without incidental, hard-wired, constraints. Here, we consider two types of hard-wired biological mechanisms: namely, range adaptation and Hebbian plasticity. We then evaluate the evidence for or against biologically-constrained ANNs using a variant of RSA (i.e., representational similarity analysis), because it exploits detailed multivariate information in the data while being robust to nuisance model misspecifications (Diedrichsen and Kriegeskorte, 2017; Diedrichsen et al., 2020; Kriegeskorte, 2008). We apply the approach to two different open fMRI datasets acquired when people make decisions under risk (Botvinik-Nezer et al., 2019). In what follows, we describe our methodological approach and evaluate its statistical properties with numerical Monte-Carlo simulations. We then report the results of the ensuing analysis of concurrent behavior and fMRI data. Finally, we discuss our results in light of the existing literature and highlight potential weaknesses and perspectives.

Methods

Biologically-constrained artificial neural networks for behavioral data

Artificial Neural Networks or ANNs provide essentially attempt to decompose a possibly complex form of information processing in terms of a combination of very simple computations performed by connected 'units', which are a mathematical abstraction of neurons. Here, we take inspiration from a growing number of studies that use ANNs as descriptive models of neural information processing, whose relative biological realism is to be gauged with neuroimaging data (Güçlü and Gerven, 2015; Kietzmann et al., 2017, 2019; Kriegeskorte and Golan, 2019).

We consider behavioral paradigms akin to decision tasks, whereby subjects need to process some (experimentally controlled) behaviourally-relevant information $u = \left\{u^{(1)}, u^{(2)}, ..., u^{(n_u)}\right\}$ to provide a response r. In what follows, we will focus on a value-based decision-making task, whereby participants have to accept or reject a risky gamble composed of a 50% chance of winning a gain G and a 50% chance of losing L, i.e., u is composed of $n_u = 2$ input features: $u = \left\{G, L\right\}$. In brief, we assume that people's behavioral response y is the output of a neural network that processes the input, i.e.: $r \approx g_{ANN}\left(u, \vartheta\right)$, where ϑ are unknown ANN parameters and $g_{ANN}\left(\bullet\right)$ is the ANN's input-output transformation function. So-called "shallow" ANNs effectively reduce $g_{ANN}\left(\bullet\right)$ to a combination of neural units organized in a single hidden layer. Here, we rather rely on ANNs with two hidden layers. As will be more apparent below, this will facilitate the introduction of Hebbian plasticity mechanisms/constraints.

We assume that each input feature $u_t^{(i)}$ is encoded into the activity of neurons $\left[x_t^{(i,1)}, x_t^{(i,2)}, ..., x_t^{(i,j)}, ..., x_t^{(i,n_x)}\right]$ of its dedicated "input layer", where n_x is the number of input neurons per input. What we mean here is that the neuron j in the input layer i responds to $u_t^{(i)}$ as follows:

$$x_t^{(i,j)} = f_1\left(u_t^{(i)}, \theta^{(i,j)}\right) \tag{1}$$

where $f_1(\bullet)$ is the activation function of neural units that compose the ANN's input layer. Collectively, the activity vector $\left[x_t^{(i,j)}\right]_{j=1,\ldots,n_x}$ forms a representation of the input $u_t^{(i)}$ in the form of a population code.

Critically, we consider activation functions that are bounded, i.e., either a sigmoid or a pseudo-gaussian mapping of inputs (see below):

$$f\left(u,\theta\right) = \begin{cases} f_{Gauss}\left(u,\theta\right) \triangleq \exp\left(-\dfrac{\left(u-\mu\right)^2}{\sigma^2}\right) \\ or \\ f_{sigmoid}\left(u,\theta\right) \triangleq \dfrac{1}{1+\exp\left(\gamma\dfrac{\mu-u}{\sigma}\right)} \end{cases} \tag{2}$$

where $\gamma \approx 1.5434$ is a scaling constant that we introduce for mathematical convenience (see Appendix 1). The parameters $\theta^{(i,j)} = \left\{\mu^{(i,j)}, \sigma^{(i,j)}\right\}$ capture the idiosyncratic properties of the neuron j in the input layer i (e.g., its firing rate threshold $\mu^{(i,j)}$ and the pseudo-variance parameter $\sigma^{(i,j)}$). Note that, when inputs u fall too far away from μ (say outside a $\pm 2\sqrt{2}\sigma$ range), both these activation functions saturate, i.e., they produce non-discriminable outputs (close to 0 or 1). In other words, the pseudo-variance parameter defines the range of inputs over which units incur no information loss. As we will see below, range adaptation effectively tunes these activation functions to minimize information loss.

Then the output of the input layers is passed to the "integration layer" $\left[z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(k)}, \ldots, z_t^{(n_z)}\right]$, i.e., the neuron k of the integration layer responds to $\left[x_t^{(i,j)}\right]_{j=1,\ldots,n_x}^{i=1,\ldots,n_u}$ as follows:

$$z_t^{(k)} = f_2\left(\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{(i,j,k)} x_t^{(i,j)}, \phi^{(k)}\right) \tag{3}$$

where $C^{(i,j,k)}$ is the connection weight from the neuron j in the input layer i to the neuron k of the integration layer, and $\phi^{(k)}$ capture idiosyncratic properties of the integration neuron k. For simplicity, we restrain our analysis to $n_z = n_x$.

The behavioral response r_t at time or trial t is then read out from the integration layer as follows:

$$r_t \approx f_{sigmoid}\left(\sum_{k=1}^{n_z} W^{(k)} z_t^{(k)}, \upsilon\right) \tag{4}$$

where the $W^{(k)}$ can be thought of as connection weights to another system that would implement the decision into an action (.e.g., the motor system).

Taken together, Equation 1-3-4 define the ANN's input-output transformation function, when no further biological constraint is introduced (see below):

$$g_{ANN}^{(0)}\left(u_t, \vartheta\right) \triangleq f_{sigmoid}\left(\sum_{k=1}^{n_z} W^{(k)} f_2\left(\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{(i,j,k)} f_1\left(u_t^{(i)}, \theta^{(i,j)}\right), \phi^{(k)}\right), \upsilon\right) \tag{5}$$

where ϑ lumps all ANN parameters together, i.e.: $\vartheta \triangleq \{W, C, \theta, \phi, \upsilon\}$, and $f_{i\in\{1,2\}}$ are either gaussian or sigmoid. A schematic summary of the ANN's double-layer structure is shown in Figure 1 below.

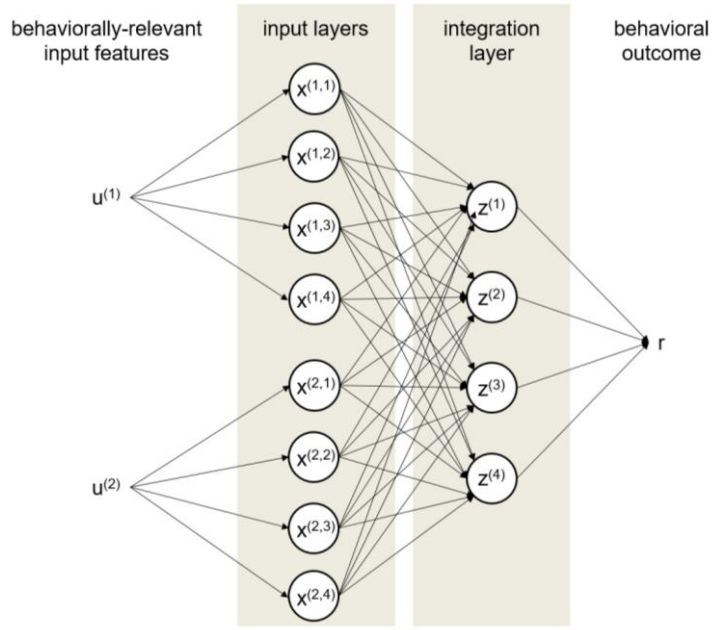

Figure 1: ***Structure of the 'default' artificial neural network.*** *Behaviorally-relevant input features first enter the 'input' layer, which then sends its multiple outputs to the 'integration' layer. Finally, a behavioral response is produced from the multiple outputs of the 'integration' layer. See the main text for mathematical notations.*

Although, strictly speaking, this ANN includes one form of biological constraint (cf. bounded units' activation functions), we will refer to it as the 'default' or 'non-constrained' ANN. Note that, provided there are enough neurons in input and integration layers, this ANN architecture can capture any value function defined on the multidimensional input space. However, it cannot capture behavioral hysteresis effects, whereby previous decisions may change the network's response to behaviorally-relevant information. This is why we now introduce range adaptation and Hebbian plasticity.

Recall that range adaptation is a mechanism by which neurons maximize the contrast of their output activity over the natural range of their inputs. Given that we used sigmoid or pseudo-gaussian activation functions (cf. Equation 2), range adaptation adaption then reduces to a learning rule on f_2 's pseudo-variance parameters, which are now time-dependent variables and seek to maximize the transmitted information, i.e., the discriminability of the outputs (see Appendix 1 for details):

$$\sigma_{t+1}^{(k)} = \sigma_t^{(k)} + \alpha_{RA} \times \left(\left| \mu^{(k)} - \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} C^{(i,j,k)} x_t^{(i,j)} \right| - \sigma_t^{(k)} \right) \quad (6)$$

where α_{RA} is the learning rate of range adaptation. Equation 6 effectively matches pseudo-variance parameters σ with the variability of the recent history of each units' inputs. In turn, units' activation functions are sampled over a range where their output activity does not saturate.

Now the two-layers structure of the ANN also enables explicit modeling of Hebbian plasticity. More precisely, the Hebbian adaption rule will strengthen the connection between input and integration units that co-vary. This recapitulates the "fire together, wire together" rule:

$$\begin{aligned} C_t^{(i,j,k)} &= c^{(i,j,k)} s\left(\kappa_t^{(i,j,k)}\right) \\ \kappa_{t+1}^{(i,j,k)} &= \kappa_t^{(i,j,k)} + \alpha_H \left(x_t^{(i,j)} z_t^{(k)} - \lambda_H \right) \end{aligned} \quad (7)$$

where $c^{(i,j,k)}$ and $\kappa_t^{(i,j,k)}$ are the static and dynamic components of between-layers connection weights, respectively, α_H is the Hebbian learning rate and λ_H is covariance threshold. Equation 7 reinforces a connection weight whenever the product of the corresponding units' outputs exceeds the threshold λ_H.

At the limit when learning rates tend to zero ($\alpha_{RA} \to 0$ or $\alpha_H \to 0$), the constrained ANNs exhibit no plastic change, i.e., they become indistinguishable from the above 'default' ANN. Otherwise, both range adaptation and Hebbian plasticity constraints make the ANN's trial-by-trial response a function of the recent history of inputs to the network. In both cases, learning rates effectively control the amount of plastic changes that modified ANNs will exhibit. Importantly, behavioral distortions and/or neural activity patterns that will be induced with these two types of plastic changes may be different. In other terms, Hebbian plasticity and range adaptation are unlikely to capture similar forms of behavioral and/or neural hysteresis effects. We will comment on the computational properties of Equations 6 and 7 in the Discussion section. Importantly, no normative model exists that can be used as a reference point to set the amount of plastic change that the decision network *should* exhibit. But

one can use observed peoples' behavioral responses to evaluate how much plastic changes the decision network actually *does* exhibit. Here, we rely on established variational Bayesian model inversion techniques to perform probabilistic parameter estimation (Daunizeau, 2017; Friston et al., 2007). To mitigate the impact of local optima, we use a twofold strategy. First, we concurrently fit the behavior trial series together with its rolling mean and variance (over a sliding temporal window whose width we set to 5 trials). Second, we use a hierarchical group-level mixed-effects approach that constrains subject-specific parameter estimates with estimated group statistics (Daunizeau, 2019). The priors on the ANNs' model parameters for the ensuing parametric 'empirical Bayes' approach are summarized in Table 1 below.

Parameter	Distributions	Rational
Pseudo-gaussian mean/ Sigmoid center	$\mu^{(i,j)} \sim \mathcal{N}\left(\frac{j}{n_x+1}, \frac{0.25}{n_x+1}\right)$	Homogenous paving of inputs
Pseudo-gaussian initial standard deviation	$\sigma_0^{(i,j)} = \lvert\theta\rvert$ with $\theta \sim \mathcal{N}\left(\frac{0.5}{n_x+1}, \frac{0.5}{n_x+1}\right)$	Overlapping pseudo-gaussian
Initial connection weights	$c^{(i,j,k)} \sim \mathcal{N}\left(\frac{1}{n_x}, \frac{1}{n_x}\right)$	Inputs averaging
Range adaptation learning rate	$\alpha_{RA} = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(-3,2)$	Gradual, stable learning
Hebbian-plasticity learning rate	$\alpha_{Hebb} = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(-3,2)$	Gradual, stable learning
Hebbian plasticity threshold	$\lambda_H = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(-1,1)$	Comparable to the average product of two bounded units
Hebbian initial strength	$\kappa_0^{(i,j,k)} = \frac{1}{1+e^{-\theta}}$ with $\theta \sim \mathcal{N}(0,0.5)$	The middle point between full and null strength

Table 1: Parameters' priors for biologically-constrained ANNs.

Note that all our behavioural analyses are performed using the VBA academic freeware (Daunizeau et al., 2014).

Assessing the neural signature of candidate biological constraints using RSA

From a statistical perspective, Equations 6 and 7 provide extra degrees of freedom when fitting the modified ANN to behavioural choices, when compared to the 'no-constraint' ANN. This means that one would expect behavior to be better explained with range adaptation and/or Hebbian constraints, irrespective of whether these constraints are realistic determinants of behavior or not. This is why it is critical to cross-validate behavioral analyses with neural data. This can be done because once fitted to behavioral data our modified ANN models make specific trial-by-trial predictions of neural activity patterns $\{x_t, z_t\}$ that can be compared to multivariate neural signals. Here, we have chosen to rely on a modified representational similarity analysis (Kriegeskorte, 2008), which possesses the following properties:

- It is simple (at least from a statistical standpoint).
- It is robust to assumptions regarding the relationship between modeled and empirical neural time series. In particular, it is not confounded by nonlinearities and/or by dimensionality differences. These, in fact, are known virtues of RSA (Diedrichsen and Kriegeskorte, 2017; Friston et al., 2019).
- It extracts multivariate information from empirical neural signals that is orthogonal to linear combinations of behaviorally-relevant inputs and behavioral responses. This is necessary (i) to provide analysis results that are orthogonal to previous mass-univariate analyses, and (ii) to prevent statistical biases towards models that best explain behavioral data.

In brief, RSA consists of evaluating the statistical resemblance between model-based and data-based 'representational dissimilarity matrices' or RDMs, which we derive as follows. Let Y be the $n_y \times n_t$ multivariate time series of (modeled or empirical) neural activity, where n_y and n_t are the number of units and trials, respectively. Note that, for model-based RDMs, 'units' mean artificial elementary units in ANNs, whereas for data-based RDMs, 'units' mean either neurons (cf. electrophysiology) or voxels

(fMRI). First, we orthogonalize Y with respect to potential confounding sources of between-trial variability, i.e.: $Y \leftarrow Y\left(I_{n_t} - X^T\left(XX^T\right)^{-1}X\right)$, where X is the $n_c \times n_t$ confounds matrix. Here, the set of confounds typically include a constant term, behaviorally-relevant inputs u, and behavioral responses r. Second, we standardize neural time series by zscoring over trials. Now let D_Y be the ensuing $n_t \times n_t$ between-trials Euclidean distance matrix:

$$D_Y = \begin{bmatrix} 0 & D_Y^{2,1} & \cdots & D_Y^{1,T} \\ D_Y^{2,1} & 0 & & D_Y^{2,T} \\ \vdots & & \ddots & \vdots \\ D_Y^{T,1} & D_Y^{T,2} & \cdots & 0 \end{bmatrix} \qquad (8)$$

$$D_Y^{t,t'} = \sum_{i=1}^{n_Y} \left| Y_t^{(i)} - Y_{t'}^{(i)} \right|$$

The matrix element $D_Y^{t,t'}$ thus measures the dissimilarity of neural patterns of activity between trial t and trial t', having removed trial-by-trial variations that can be explained as linear combinations of behaviorally-relevant inputs and behavioral responses. We define the ensuing RDM as the lower-left triangular part of D_Y.

In what follows, model-based RDMs are derived using the integration layer of our modified ANNs (i.e. $Y_{ANN} = \left[z_1, z_1, ..., z_{n_t}\right]^T$), after having fitted the corresponding model parameters to behavioral responses. Data-based RDMs are derived from the fMRI time series. Here, Y_{fMRI} is obtained by deconvolving BOLD time series from the hemodynamic response function with a Dirac delta or stick basis function set that is time-locked to trial events (Dale, 1999). RSA then proceeds with the statistical comparison of $D_{Y_{ANN}}$ and $D_{Y_{fMRI}}$. In line with recent methodological developments of RSA, we first bin RDMs into 20 quantiles and then compute the Pearson correlation $\rho = corr\left(RDM_{ANN}, RDM_{fMRI}\right)$ between the binned RDMs. Group-level statistical significance of RDMs' correlations can be assessed

using one-sample t-tests on the group mean of Fischer-transformed RDM correlation coefficients ρ (see below). Figure 2 below recapitulates the ensuing ANN-RSA approach.

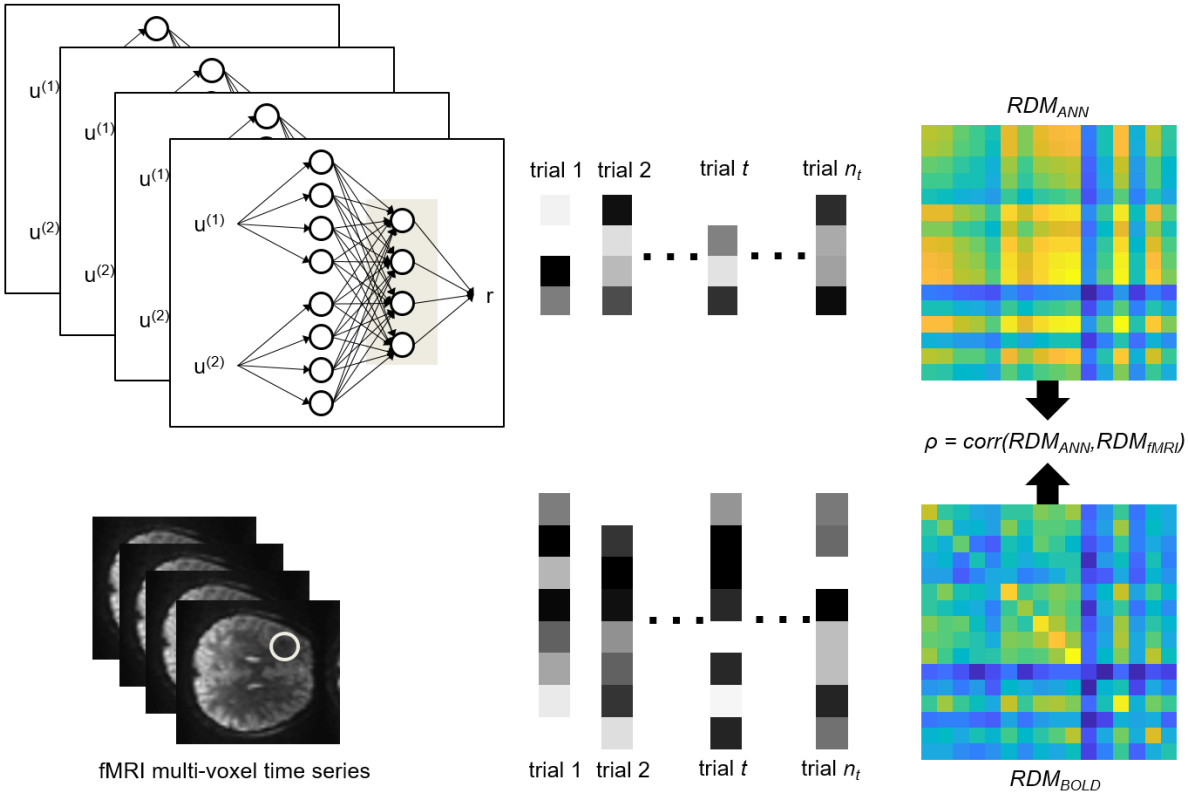

Figure 2: ***Summary data-analysis pipeline of the ANN-RSA approach.*** *First, trial-by-trial profiles of the ANN's response to behaviourally-relevant inputs (in the integration layer) are estimated. Second, corresponding trial-by-trial multivariate patterns of fMRI activity are extracted in each ROI of interest. Third, corresponding model-based and fMRI-based RDM are derived, whose correlation* ρ *serves as the RSA summary statistics (which then enters subsequent statistical significance testing).*

Note that our ANN-RSA approach does not *a priori* favor more complex ANNs (i.e., ANNs with more parameters). When fitted to behavioral data, more complex ANNs (i.e., those that include range adaptation or Hebbian plasticity) are expected to yield greater explanatory power. The RDM correlation ρ exhibits no such bias, however. This is because, once fitted to behavioural data, estimated ANN activity patterns and their ensuing RDMs have no degree of freedom whatsoever. In particular, this means that default (non-constrained) ANNs may show a greater RDM correlation than ANNs that include range adaptation or Hebbian plasticity. In turn, this enables a simple statistical procedure for comparing candidate models based on group-level comparisons of RDM correlations (see below).

Note on statistical testing and model comparison

Recall that our model space is factorial, with two orthogonal modeling factors: (i) our factor of interest has three 'levels': no constraint, range adaptation or Hebbian plasticity, and (ii) our factor of no interest has two 'levels': sigmoid versus pseudo-gaussian neural activation functions. This means that we will be comparing 2x3=6 models. When assessing the statistical significance of the ensuing model comparison, we will be using a variant of composite null testing. Let $p_m^{m'}$ be the p-value associated with the elementary pairwise comparison of model m and m', whose null hypothesis is $H_0^{(m,m')}: \rho_m \leq \rho_{m'}$, where ρ_m is the corresponding Fisher-transformed RDM correlation ($p_m^{m'}$ can be evaluated using paired t-tests on RDM correlations). For each model $m \in [1,6]$, we ask whether its RDM correlation is the highest among the candidate models. This induces the following composite null hypothesis: $H_0^{(m)}: \rho_m \neq \max_{m'} \rho_{m'}$. The maximum p-value statistics $\hat{p}_m = \max_{m'} p_m^{m'}$ yields a valid test of the composite null hypothesis, though not necessarily maximally efficient (Wasserman, 2004). Because $H_0^{(m)}$ is the conjunction of elementary pairwise null hypotheses $H_0^{(m,m')}$, we refer to this approach as "conjunctive null testing".

One may also want to evaluate the statistical significance of the comparison of RDM correlations across levels of our factor of interest, irrespective of our factor of no interest. The corresponding null hypothesis involves a disjunctive/conjunctive combination of elementary null hypotheses. For example, if one wants to test whether range adaptation has a significantly higher RDM correlation than Hebbian or default (no-constraint) ANNs, the corresponding null hypothesis $H_0^{(RA)}$ is defined as:

$$H_0^{(RA)}: \begin{cases} \rho_{RA,Gauss} \neq \max\limits_{m' \neq \{RA,sigmoid\}} \rho_{m'} \\ AND \\ \rho_{RA,sigmoid} \neq \max\limits_{m' \neq \{RA,Gauss\}} \rho_{m'} \end{cases} \quad (9)$$

The following p-value then yields a valid statistical test of $H_0^{(RA)}$:

$$\widehat{p}_{RA} = 2 \times \min\left[\max_{m' \neq \{RA, sigmoid\}} p_m^{m'}, \max_{m' \neq \{RA, Gauss\}} p_m^{m'} \right] \quad (10)$$

By design, the ensuing “disjunctive/conjunctive” approach cannot conclude about the underlying activation functions, i..e. it does not discriminate between sigmoid and pseudo-gaussian functional forms. However, it pools evidence over levels of our factor of no interest, which eventually improves statistical power. This is a frequentist -and simpler- variant of so-called "family inference" in Bayesian model comparison (Penny et al., 2010), where one marginalizes over modeling factors of no interest, effectively trading statistical power against inference resolution. We will see a direct demonstration of the disjunctive/conjunctive approach below.

fMRI study of risk attitudes: experimental design

In this work, we compare the neural evidence for candidate biological constraints (range adaptation versus Hebbian plasticity) on behaviorally-relevant neural information processing using a re-analysis of the NARPS dataset (Botvinik-nezer et al., 2019), openly available on openneuro.org (Poldrack et al., 2013). This dataset includes two studies, each of which is composed of a group of 54 participants who make a series of decisions made of 256 risky gambles. On each trial, a gamble was presented, entailing a 50/50 chance of gaining an amount G of money or losing an amount L. As in Tom et al. (2007), participants were asked to evaluate whether or not they would like to play each of the gambles presented to them (strongly accept, weakly accept, weakly reject or strongly reject). They were told that, at the end of the experiment, four trials would be selected at random: for those trials in which they had accepted the corresponding gamble, the outcome would be decided with a coin toss, and for the other ones -if any-, the gamble would not be played. In the first study (hereafter: "equal range"

group), participants decided on gambles made of gain and loss levels that were sampled from the same range (G and L varied between 5 and 20 $). In the second study (hereafter: the "equal indifference" group), gain levels scaled to double the loss levels (L varied between 5 and 20$, and G varied between 10 and 40$). In both studies, all 256 possible combinations of gains and losses were presented across trials, which were separated by 7 seconds on average (min 6, max 10).

MRI scanning was performed on a 3T Siemens Prisma scanner. High-resolution T1w structural images were acquired using a magnetization prepared rapid gradient echo (MPRAGE) pulse sequence with the following parameters: TR = 2530 ms, TE = 2.99 ms, FA = 7, FOV = 224 × 224 mm, resolution = 1 × 1 × 1 mm. Whole-brain fMRI data were acquired using echo-planar imaging with multi-band acceleration factor of 4 and parallel imaging factor (iPAT) of 2, TR = 1000 ms, TE = 30 ms, flip angle = 68 degrees, in-plane resolution of 2X2 mm 30 degrees of the anterior commissure-posterior commissure line to reduce the frontal signal dropout, with a slice thickness of 2 mm and a gap of 0.4 mm between slices to cover the entire brain. See https://www.narps.info/analysis.html#protocol for more details. Data preprocessing included standard realignment and movement correction steps. Note that we excluded 5 participants from the 'equal-range' group because the misalignment between functional and anatomical scans could not be corrected. No spatial smoothing was applied.

Previous mass-univariate analyses of these datasets, including a recent study of the analysis variability among multiple research groups (Botvinik-nezer et al., 2019), provided evidence for the implication of multiple brain systems in response to either gains and/or losses, in particular: the ventromedial prefrontal cortex or vmPFC, the dorsolateral prefrontal cortex or dlPFC, the anterior cingulate cortex or ACC, the posterior cingulate cortex or PCC, the Amygdala, the Striatum and the Insula. Given the anatomo-functional variability of these regions, we opted for a multiple ROI analysis. Using the NeuroQuery website (Dockès et al., 2020), we selected spatial maps based on the following 12 terms: vmPFC, dlPFC, ACC, dACC, PCC, Amygdala, Striatum, and Insula. We also included primary motor and primary visual cortices, which serve as sensory/motor control regions. Then we took the 2000-th

strongest voxels, excluded those that belonged to clusters smaller than 200 voxels, smooth the resulting map, filter out white matter overlaps, and kept the 200 strongest voxels of each remaining clusters. This procedure yielded 18 approximately spherical ROIs spanning both hemispheres, which are shown in Figure 3 below.

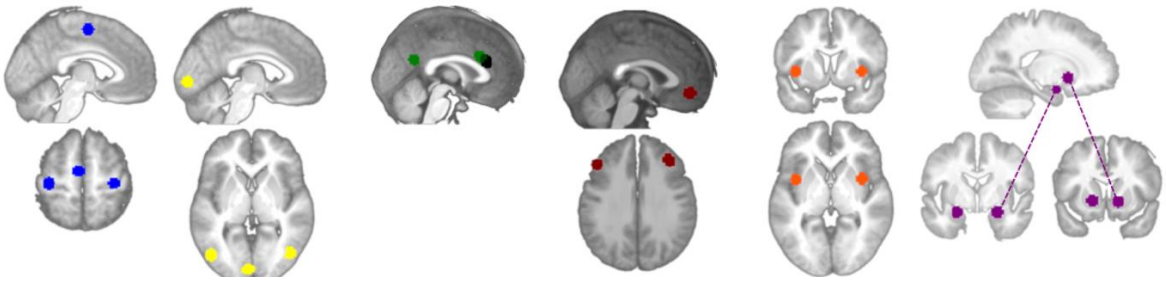

Figure 3: ***Regions of interest.*** *Control ROIs: Motor left, median, and right (blue), Visual left, median, right (yellow). ROIs of interest: PCC, ACC and dACC (green), vmPFC and dlPFC left and right (red), Insula left and right (orange), Amygdala left and right, and Striatum left and right (purple).*

In each ROI, we regressed trial-by-trial activations with SPM through a GLM that included one stick regressor for each trial (at the time of the gamble presentation onset), which was convolved with the canonical HRF. To account for variations in hemodynamic delays, we added the basis function set induced by the HRF temporal derivative (Hopfinger et al., 2000). To correct for movement artifacts, we also included the six head movement regressors and their squared values. We then extracted the 256 trial-wise regression coefficients in each voxel of each ROI. Finally, we orthogonalized the resulting fMRI trial series w.r.t. gains, losses, and choices, zscored them and computed the 18 ROI-specific RDMs.

Results

Assessing expected model confusion using numerical Monte-Carlo simulations

Prior to presenting our fMRI analyses, we ought to provide evidence that our combined ANN-RSA approach exhibits the statistical robustness that is required for a reliable interpretation of results. In particular, one may ask whether the approach is robust to modeling assumptions regarding (i) the (necessarily underestimated) dimensionality of ANNs that process behaviorally-relevant information, and (ii) the form of units' activation functions (cf. sigmoid versus pseudo-gaussian). More precisely, we ask whether the approach discriminates between the three candidate biological mechanisms of interest (range adaptation, Hebbian plasticity, and 'default'), even when the data are generated with higher-dimensional ANNs. We thus performed a series of Monte-Carlo simulations that recapitulates the design of the fMRI experiment.

We considered a decision task that requires the integration of two inputs $u = \left\{u^{(1)}, u^{(2)}\right\}$ that vary randomly across 256 trials. We simulated six series of datasets, corresponding to the 2x3=6 alternative modified ANN models described above. Each dataset was composed of 20 virtual subjects whose trial-by-trial behavior and neural responses were generated under an ANN with sets of either $n_x = 20$, 30, or 50 neural units. We allowed for inter-individual variability, derived from sampling ANN parameters under their respective prior probability density functions (cf. Table 1). Each simulated dataset was then analyzed using the ANN-RSA approach described above. In brief, each behavioral trial series was fitted with the 2x3 candidate ANNs, and the resulting estimated neural activity profiles were compared to simulated neural activity profiles using our modified RSA. Importantly, fitted ANNs contained smaller sets of $n_x = 10$ units. For each dataset, we then compared models using conjunctive null testing. We repeat this procedure 50 times and keep track of all positive tests (with a 5% significance level). The

upper panel of Figure 4 shows the frequency of positive conjunctive testing for all candidate models for each type of simulated data.

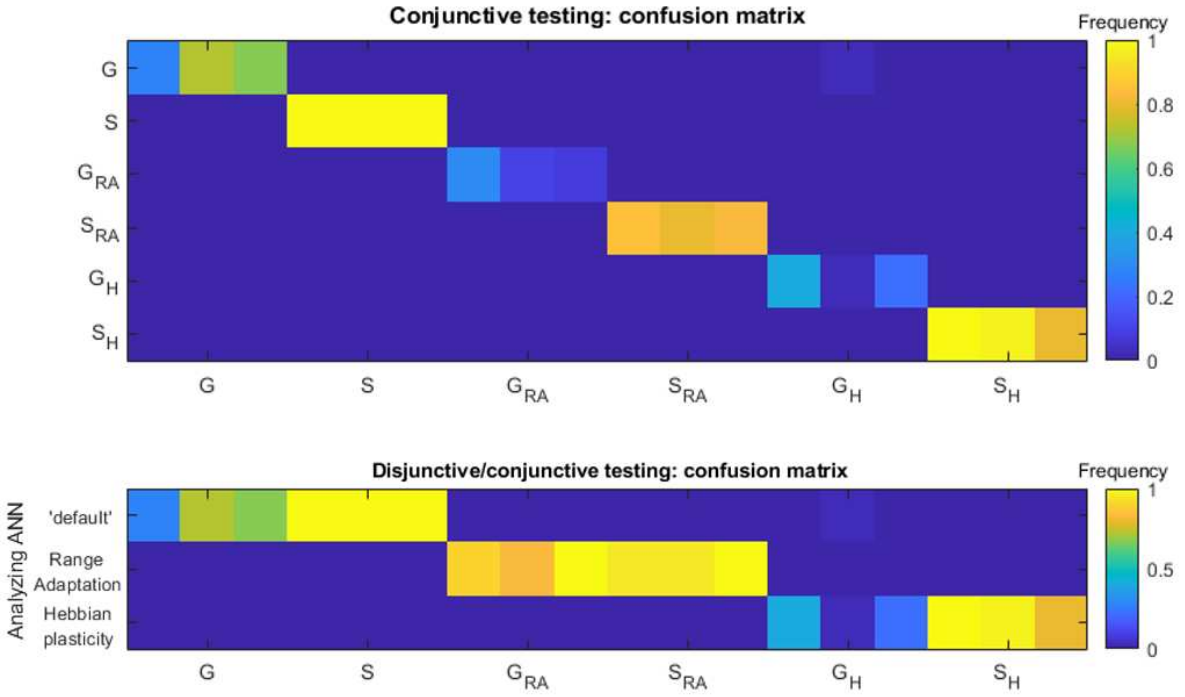

Figure 4: ***Robustness of the ANN-RSA approach: Monte-Carlo simulations.*** *In what follows, so-called "generative" ANNs were used to simulate data. They can be of 2x3=6 sorts: pseudo-Gaussian/sigmoid 'default' ANNs, pseudo-Gaussian/sigmoid range adaptation ANNs, and pseudo-Gaussian/sigmoid Hebbian ANNs. Each of these sorts of ANNs had three possible dimensions with sets of* $n_x = 20, 30,$ *or* 50 *units. In contrast, "analyzing" ANNs only included sets of* $n_x = 10$ *units. Upper panel: confusion matrix of the conjunctive testing approach. The rate at which each "analyzing" ANN (y-axis) exhibits significantly higher RDM correlations than other models, for each "generative" ANNs (x-axis) is color-coded. The three alternative dimensions of "generative" ANNs are presented side to side, from left to right. Lower panel: confusion matrix of the disjunctive/conjunctive approach. Same format, except that the y-axis now shows candidate mechanisms.*

First, note that the conjunctive approach exhibits almost no model confusion. More precisely, the maximum frequency of a model selection error is about 10% (generative ANN = pseudo-gaussian ANN with Hebbian plasticity and 30 units, analyzing ANN= pseudo-gaussian ANN). However, its statistical power is variable (from about 92% ±2% on average for all sigmoid ANNs to about 31% ±25% on average for all pseudo-gaussian ANNs). In other words, the conjunctive testing approach may be too conservative in detecting the correct ANN. Second, the dimensionality of generative ANNs seems to have almost no impact on statistical power. In other words, the relatively small dimensionality of analyzing ANNs (when compared to generative ANNs) does not seem to impair the method's ability to detect the correct underlying mechanism.

Now the lower panel of Figure 4 shows the frequency of positive disjunctive/conjunctive testing for the three types of biological mechanisms (no constraint, range adaptation, or Hebbian plasticity) for each type of simulated data. One can see that model confusion is similar to the conjunctive approach above. However, statistical power is much improved, in particular for detecting range adaptation (94% ±7% on average). Here again, the dimensionality of generative ANNs seems to have no impact on statistical power.

In conclusion, the ANN-RSA approach is robust to violations of modeling and statistical assumptions, including the low dimensionality of analyzing ANNs or the distribution of test statistics. In particular, this implies that, if a candidate mechanism eventually reaches statistical significance using the disjunctive/conjunctive approach, then we can safely infer that it is a more likely explanation of fMRI activity patterns than other candidate mechanisms.

Behavioural analyses

Each participant's choice sequence data were fitted with the six candidates ANNs, as well as with a simple logistic model. We used sets of $n_x = 4$ units and normalized the gain and loss levels by their averaged sum before feeding them to the input layer. The latter logistic model is the typical agnostic modeling choice in decision paradigms of this kind and was used to measure loss aversion in a previous study relying on the same behavioral design (Tom et al., 2007). Here, it will serve as a reference model for evaluating the predictive power of ANNs. Each group was fitted independently through the VBA empirical Bayes procedure. All summary statistics of these behavioural analyses are provided in Tables 1 and 2 of the Appendix. Figure 5 below summarizes the fit accuracy of the seven models for the 'equal range' group.

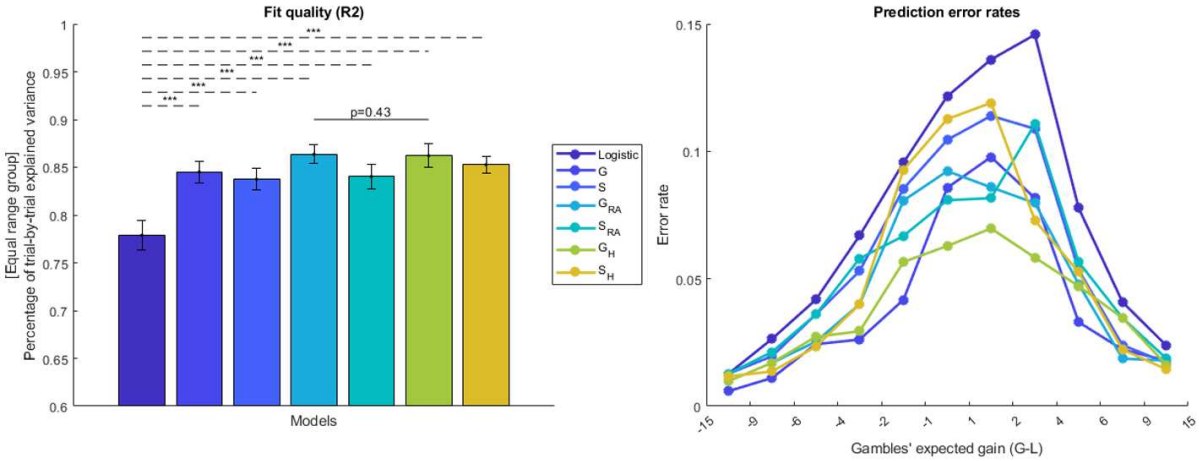

Figure 5: ***Behavioral results: 'equal range' group.*** *Left panel: mean percentage of variance explained in trial-by-trial choices ± one standard error of the mean (y-axis) is shown for each candidate model (x-axis: from left to right: logistic reference model, pseudo-Gaussian default ANN, sigmoid default ANN, pseudo-Gaussian range adaptation ANN, sigmoid range adaptation ANN, pseudo-Gaussian Hebbian ANN and sigmoid Hebbian ANN). Right panel: The average rate of prediction error (y-axis) is plotted as a function of gambles' expected gain (i.e., G-L, x-axis) for each candidate model (same color code as left panel). Note that the indifference point (maximal prediction error) seems to be biased towards positive expected gains.*

First, one can see that all candidate ANNs perform much better than the simple (reference) logistic model. In fact, they all exhibit a significantly higher percentage of explained variance (all $p<10^{-5}$). It turns out that most of the fit improvement lies around the indifference point, where gains and losses balance out (cf. right panel of Figure 4). Around that point (i.e., within the [-1,4] interval of expected utility), the logistic reference model necessarily makes unreliable predictions and yields an average error rate of about 12.2% to 14.6%. In comparison, ANNs seem to be able to reduce the apparent randomness in participants' choices, even around the indifference point. This is clearly the case for the model that achieves the lowest average error rate (about 6.3% to 7.0%): namely: the 'pseudo-gaussian Hebbian' ANN. A likely explanation here is that Hebbian plasticity may effectively change, in a deterministic but nonlinear manner, the network response to repetitions of -otherwise indifferent- gambles. In turn, seemingly random choices may be, at least partially, predicted from the history of past network inputs. This may be taken as evidence against the range adaptation mechanism, which exploits qualitatively similar history-dependent effects to find predictors of peoples' choices around the indifference point. However, it is difficult to conclude from behavioral data alone, because there is no strong statistical evidence that the 'pseudo-gaussian Hebbian' ANN has better explanatory power

than the 'pseudo-gaussian range adaptation' ANN, which is the next best model in terms of behavioral fit accuracy (average R2 difference = 0.1% ± 5.6%, p=0.43).

Figure 6 below presents the results of the same analysis for the 'equal indifference' group.

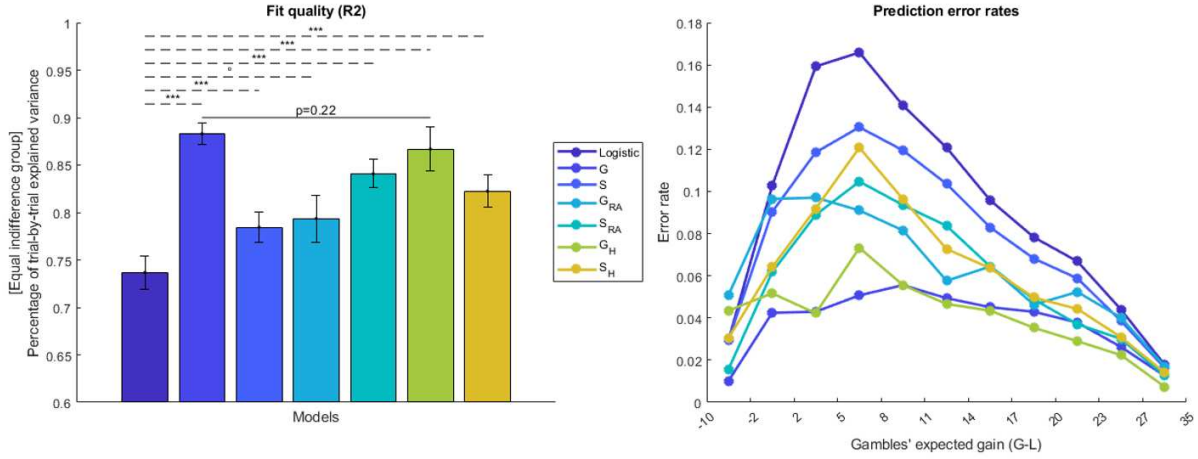

Figure 6: ***Behavioral results: 'equal indifference' group.*** *Same format as Figure 5.*

In brief, the same observations can be made, i.e., the behavioral analysis replicates on this second study. In particular, here again, the 'pseudo-gaussian Hebbian' ANN achieves an average error rate of about 6.4% to 9.2% around the indifference point but shows no significant difference in explanatory power with the next best model (average R2 difference = 1.6% ± 15.1%, p=0.22).

At this stage, one would conclude that although biologically-constrained ANNs seem to provide clear improvements over simple statistical behavioural models, behavioral data alone does not clearly discriminate between candidate underlying biological mechanisms/constraints.

FMRI analyses

We now aim at identifying the neural signature of candidate biological mechanisms/constraints that may determine people's choice sequences.

To begin with, we simply ask whether *any* candidate model actually explain multivariate fMRI time series in *any* ROI that we included in our analysis. Figure 7 below summarizes the ANN-RSA analysis, in terms of the group-average RDM correlations ρ for each pair of candidate model and ROI ('equal range' group). Table 3 in the Appendix provides the ensuing p-value of RDM correlations' group-level statistical significance (H0: $\rho \leq 0$, one-sided t-test). Note that instead of using units activity, we computed the RDM of the logistic model from the gain and loss levels weighted by the regression coefficients, and orthogonalized from the subject's choices only.

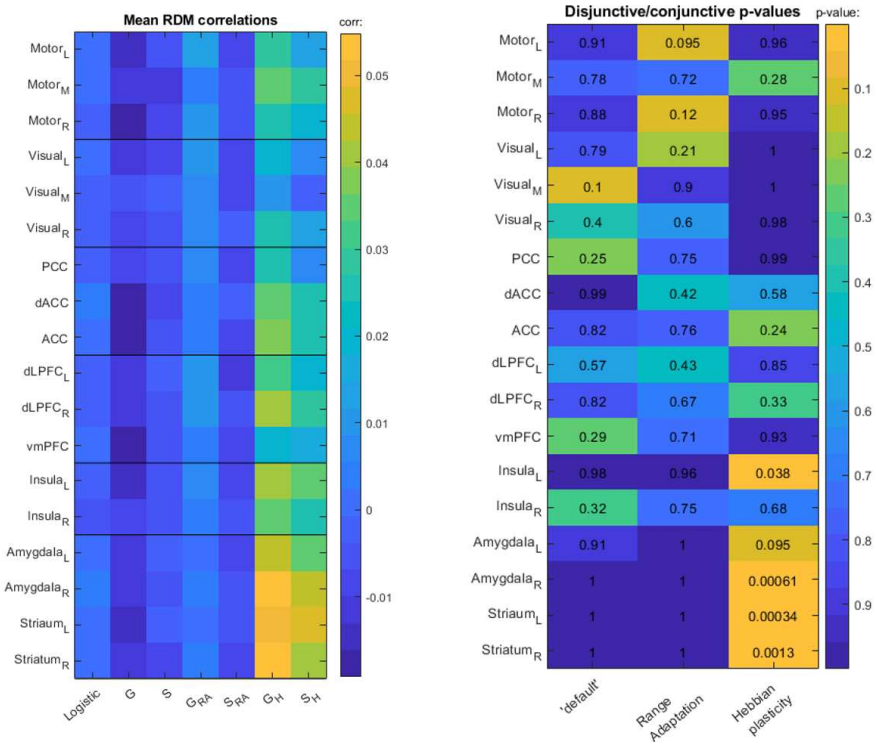

Figure 7: ***FMRI results: 'equal range' group.*** *Left: group means RDM correlations are shown for each candidate model (x-axis, from left to right: logistic reference model, pseudo-Gaussian default ANN, sigmoid default ANN, pseudo-Gaussian range adaptation ANN, sigmoid range adaptation ANN, pseudo-Gaussian Hebbian ANN, and sigmoid Hebbian ANN) and each ROI (y-axis, from top to bottom: left motor, medial motor, right motor, left visual, medial visual, right visual, PCC, dorsal ACC, ACC, left DLPFC, right DLPFC, vmPFC, left Insula, right Insula, left Amygdala, right Amygdala, left ventral Striatum, right ventral Striatum). Right: group-level p-values of the disjunctive/conjunctive approach to comparing candidate mechanisms are shown for each mechanism (x-axis, for left to right: 'default', range adaptation, and Hebbian plasticity) and each ROI (y-axis, same order as left panel).*

One can see that non-Hebbian models exhibit very small RDM correlations when compared to Hebbian models. Also, the RDM correlations of all models (including Hebbian models) are very weak in control (visual and motor) ROIs. More precisely, no model reaches statistical significance in control regions when correcting for multiple comparisons (all p>0.0008, Bonferroni-corrected threshold=0.00046). In

fact, only RDM correlations of Hebbian ANNs reach statistical significance, and only in right DLPFC (pseudo-Gaussian: p=0.0004, sigmoid: p=0.0004), left insula (pseudo-Gaussian: p=0.0004, sigmoid: $p<10^{-4}$), left amygdala (pseudo-Gaussian trend: p=0.0006, sigmoid: p=0.0001), right amygdala (pseudo-Gaussian: $p<10^{-4}$, sigmoid: $p<10^{-4}$), left striatum (pseudo-Gaussian: $p<10^{-4}$, sigmoid: $p<10^{-4}$) and right striatum (pseudo-Gaussian: p=0.0001, sigmoid: p=0.0002).

We then compared Hebbian plasticity to other biological mechanisms of interest using disjunctive/conjunctive testing, whose ensuing p-values are shown in Figure 7 (right panel Bonferroni-corrected threshold=0.0028). We found that the comparison of RDM correlations reached statistical significance in bilateral Striatum (left Striatum: p=0.0003, right Striatum: p=0.001) and in the right Amygdala (p=0.0006). In control ROIs, no comparison of RDM correlations achieves statistical significance (all p>0.1, uncorrected). Furthermore, the RDM correlations of range adaptation are never significantly higher than those of other models (all p>0.095, uncorrected).

Figure 8 below summarizes the results of the same analysis for the 'equal indifference' group (Table 4 in the Appendix provides the ensuing p-value of RDM correlations).

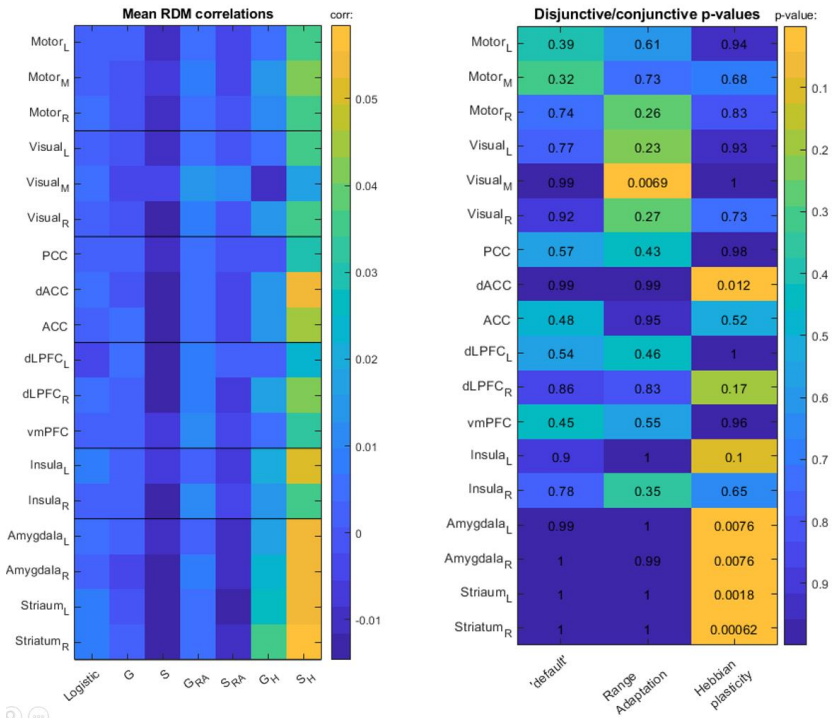

Figure 8: ***FMRI results: 'equal indifference' group.*** *Same format as Figure 7.*

In brief, results remarkably replicate the 'equal range' study. Here again, the RDM Hebbian plasticity reaches statistical significance in the left Striatum (pseudo-Gaussian: p=0.0004, sigmoid: $p<10^{-4}$), right Striatum (pseudo-Gaussian: p=0.0001, sigmoid: $p<10^{-4}$), left amygdala (pseudo-Gaussian trend: p=0.002, sigmoid: $p<10^{-4}$) and right Amygdala (pseudo-Gaussian: p=0.0002, sigmoid: $p<10^{-4}$). We note that here, the RDM correlations of sigmoid-Hebbian ANNs reach statistical significance in all other ROIs except in the medial visual cortex (all $p<10^{-4}$). Notably, the RDM correlations of Hebbian ANNs are only significantly higher than other mechanisms of interest in bilateral Striatum (left Striatum: p=0.0018, right Striatum: p=0.0006). However, there is a trend in bilateral Amygdala (left/right Amygdala: p=0.0076). In control ROIs, no model comparison achieves statistical significance, and the RDM correlations of range adaptation are never statistically higher than those of other mechanisms.

At this stage, one may safely conclude that Hebbian plasticity is a more likely explanation for fMRI activity patterns during risky decisions than range adaptation (or the default, non-constrained, biological scenario). But is Hebbian plasticity impairing or enabling adaptive behavior? Numerical

simulations on fitted Hebbian ANNs show that reducing Hebbian learning rates α_H (keeping all other estimated parameters the same) altered the decisions' sensitivity to small gains and high losses, effectively increasing loss aversion. But computational investigations of this sort cannot tell us whether and how people's behavior change when their brain activity displays more *Hebbian-ness*, i.e., when it becomes more similar to predictions from Hebbian ANNs. We thus ask whether inter-individual differences in *Hebbian-ness* may explain inter-individual differences in behavior, in particular: *choice inconsistency*. We define the *Hebbian-ness* of fMRI activity patterns in terms of the increase in neural evidence for the Hebbian ANN when compared to the default (non-constrained) ANN. Let R_m^2 be the percentage of explained variance in the fMRI RDM using the model m (in each ROI). We then measure *Hebbian-ness* using the following pseudo F-score: $R_{Hebb}^2 - R_{default}^2$. We define *choice inconsistency* in terms of the number of choices that contradict the logistic reference model, once it has been fitted to behavioral data. This effectively measures the rate of decisions, close to a subject's subjective indifference point, that contradicts its average preference. We then regress choice inconsistency against *Hebbian-ness* in bilateral Striatum and Amygdala concurrently (independently for both sigmoid and pseudo-gaussian ANNs). Figure 9 below summarizes this analysis for both groups of participants.

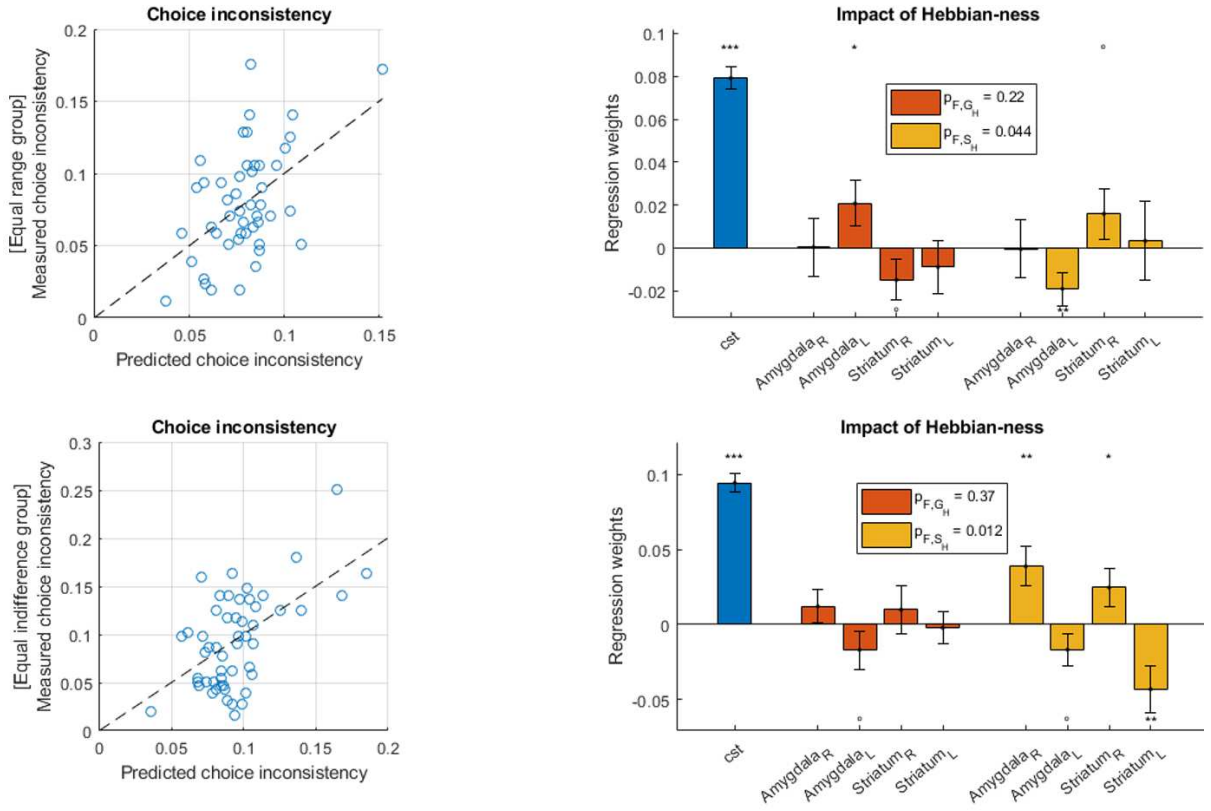

Figure 9: ***Analysis of inter-individual variability.*** *Upper-left panel: measured (x-axis) and predicted (y-axis) rates of choice inconsistency are plotted against each other for the 'equal range' group (each dot is a participant). Lower-left panel: same as above, for the 'equal indifference' group. Upper-right panel: the normalized regression weight estimates (y-axis) are shown for each corresponding ROI (x-axis, from left to right: left Amygdala, right Amygdala, left ventral Striatum, right ventral Striatum), for both pseudo-gaussian (red) and sigmoid (yellow) Hebbian ANNs. Lower-right panel: same as above, for the 'equal indifference' group.*

One can see that, when using pseudo-gaussian ANNs, *Hebbian-ness* does not predict inter-individual differences in choice inconsistency ('equal range' group: p=0.22, 'equal indifference' group: p=0.37, omnibus F-test). However, when using sigmoid ANNs, inter-individual differences in choice inconsistency can be predicted from fMRI measures of *Hebbian-ness* ('equal range' group: p=0.044, 'equal indifference' group: p=0.012, omnibus F-test). Now whether *Hebbian-ness* facilitates or hinders choice consistency seems to depend upon where in the brain it is measured. More precisely, increasing *Hebbian-ness* in the left amygdala decreases choice inconsistency ('equal range' group: p=0.019, 'equal indifference' group trend: p=0.057), whereas (right-)striatal *Hebbian-ness* increases it ('equal range' group trend: p=0.18, 'equal indifference' group: p=0.026). We note that *Hebbian-ness* in the right Amygdala and left Striatum does not seem to have a robust effect on choice inconsistency, since

statistical significance is reached only for the 'equal indifference' group (right Amygdala: p=0025, left Striatum: p=0.0036), but not for the 'equal range' group (right amygdala: p=0.97, left Striatum: p=0.85).

Discussion

In this work, we identify the neural signature of candidate biological constraints and/or mechanisms that may shape or distort neural information processing. Rather than using normative models of behavior, we quantify the (potentially idiosyncratic) impact of biological constraints by fitting constrained ANNs to people's behavioral responses. We then use RSA to compare the estimated neural activity profiles to multivariate fMRI signals. Using numerical Monte-Carlo simulations, we demonstrate that the ensuing ANN-RSA approach is robust to modeling and statistical assumptions of no interest. We then show, on two independent fMRI studies, that (i) seemingly indifferent choices in risky gambles are partially determined by range adaptation and/or Hebbian plasticity, (ii) multivariate activity in Striatum and Amygdala during choice is better explained by Hebbian plasticity than with range adaptation, and (iii) the *Hebbian-ness* of striatum and amygdala activity profiles predicts inter-individual differences in choice inconsistency.

From a methodological standpoint, our main contribution is to show how to quantify the neural evidence for or against incidental, hard-wired, biological constraints on behaviorally-relevant information processing. With this aim, we retain the simplicity of established 'model-based' fMRI approaches (Borst et al., 2011; O'Doherty et al., 2007), which proceed by cross-validating the identification of hidden computational determinants of behavior with neural data. In addition, we leverage the flexibility of ANNs and RSA to extend the breadth of empirical questions that can be addressed using dual computational/behavioural means.

In particular, this enables us to quantify the statistical evidence for neurophysiological mechanisms that are difficult –if not impossible- to include in computational models that are defined at Marr's *algorithmic* level (McClamrock, 1991), e.g., normative models of behavior (as derived from, e.g., learning or decision theories) and/or cognitive extensions thereof. Hebbian plasticity is a paradigmatic example of what we mean here. Recall that it was initially proposed as an explanation -at the neural or Marr's *implementational* level- for learning, memory, and sensory adaptation (Hebb, 1950). Since

then, Hebbian-like synaptic plasticity that serves well-defined computational purposes of this sort has been superseded by theoretical frameworks that transcend the three Marr's analysis levels, e.g., the "Bayesian brain" hypothesis (Aitchison and Lengyel, 2017; Doya et al., 2007; Friston, 2012). But hard-wired biological mechanisms of this sort may not always be instrumental to the cognitive process of interest. In turn, it may be challenging to account for incidental biological disturbances of neural information processing, when described at the algorithmic level. A possibility here is to conceive of these disturbances as some form of random noise that perturbs cognitive computations (Drugowitsch et al., 2016; Wyart and Koechlin, 2016). That these stochastic scenarios remain agnostic about the underlying (most likely hard-wired and deterministic) biological processes is both their strength and their weakness.

Of course, the field has been using neural network models of behavior for decades (Deco et al., 2013; Frank, 2006; Jocham et al., 2012; Rigoux and Daunizeau, 2015; Wang, 2008). However, existing models are typically difficult to generalize beyond the empirical frame within which they have been derived. This is because model-based predictions typically rely on many assumptions that are specific to the neural circuit and/or the cognitive process of interest. In contrast, we take inspiration from recent theoretical work promoting the advantages of pairing ANNs with RSA (Kriegeskorte and Diedrichsen, 2016, 2019), and search for neural evidence buried in multivariate patterns of brain activity while marginalizing over modeling assumptions of no interest. The aim here is to keep the modeling simple and protect the ensuing statistical inference from quantitative assumptions that have no theoretical or empirical support (cf., e.g., ANN dimensionality and/or sigmoid versus pseudo-gaussian activation functions). Although the numerical simulations we present here tend to validate our statistical treatment, we think that this kind of problem is more flexibly solved using the so-called 'family inference' in the context of Bayesian model comparison (Penny et al., 2010). In brief, the family inference is an optimal method for pooling statistical evidence over modeling factors of no interest and has proven both specific and sensitive in the context of large model spaces (Penny and Ridgway,

2013). This would be most likely needed when extending the set of candidate biological constraints and/or when studying their interactions (see below).

A related point is the issue of defining which data feature(s) is eventually compared to model predictions. By construction, RSA assumes that candidate scenarios can be faithfully evaluated in terms of their ability to predict the trial-to-trial similarity/dissimilarity of multivariate (fMRI) patterns of neural activity. At the very least, this discards potentially relevant information, e.g., voxels' spatial location and peri-stimulus dynamics are lost (Kriegeskorte and Diedrichsen, 2016). Whether and how one may improve the statistical efficiency and robustness of RSA are unresolved issues (Diedrichsen and Kriegeskorte, 2017; Diedrichsen et al., 2020; Friston et al., 2019; Kriegeskorte and Diedrichsen, 2019). In our context, this has two practical consequences. First, we used control sensory and motor ROIs to demonstrate the anatomo-functional specificity of our inference. Problematic here is the fact that we relied on negative results (in control ROIs), which may follow from the limited statistical efficiency of RSA. In the context of classical mass-univariate approaches, the issue of comparing different brain regions is known to be bound to many intricate confounds (Henson, 2006). How these interact with the statistical properties of RSA is virtually unknown. Second, one may question the way we defined our set of confounds when deriving the ANN and fMRI RDMs. More precisely, we removed trial-by-trial variations that can be explained by linear combinations of inputs and outputs. This is important if one is to (i) draw inferences that are orthogonal to linear univariate event-related fMRI analyses, and (ii) prevent a bias towards models that fit behavior best. Note that the latter issue is critical for our definition of *Hebbian-ness*, whose inter-individual variations may otherwise be driven by statistical artifacts that grow with behavioral atypicality. The obvious cost of this conservative strategy is in terms of information loss. Although our results are qualitatively unchanged when excluding inputs and outputs from the set of confounds (not shown), this may not always be the case. In our opinion, addressing these sorts of issues may require the development of more sophisticated computational approaches that can treat behavioral and neural data in a statistically symmetrical

manner (Rigoux and Daunizeau, 2015; Turner et al., 2013, 2016, 2019). We intend to pursue this type of approach in subsequent publications.

At this point, and given the above limitations, we acknowledge that our neuroscientific claim is quite modest. In brief, our results support Hebbian plasticity as a valid alternative to range adaptation in the context of risky gambles. That we eventually identify the Striatum and the Amygdala to be specifically involved in this context is well aligned with the existing literature. On the one hand, the ventral Striatum is known to encode value and risk (Schultz et al., 2008), and the tendency to opt for a risky choice increases with the magnitude of the striatal response to risk (Christopoulos et al., 2009; Kuhnen and Knutson, 2005). In fact, the same experimental protocol as we use here ('equal indifference' range) already served to demonstrate that the differential striatal responses to losses and gains drive inter-individual variations in loss aversion (Tom et al., 2007). On the other hand, it was also shown that the prospect of a possible loss might activate Amygdala, which would trigger a cautionary brake on behavior that facilitates loss aversion (Martino et al., 2006, 2010). How ventral Striatum and Amygdala eventually interact with each other to determine loss aversion is unknown, and the present study does not resolve this debate. In line with recent studies of hysteretic effects in the brain's decision system (Conen and Padoa-Schioppa, 2019; Rangel and Clithero, 2012; Soltani et al., 2012), we rather focus on seemingly indifferent and/or inconsistent choices, which remain otherwise unexplained. The present results illustrate how neuroimaging can be used to directly test whether candidate hard-wired, incidental, biological constraints may impact on behavior: in this case, the hysteretic effects of range adaptation and/or Hebbian plasticity. Hebbian plasticity, but not range adaptation, was observed in both brain systems that were previously shown to regulate loss aversion. Retrospectively, however, many other candidate mechanisms may, in principle, explain such hysteretic effects, e.g., homeostatic plasticity (Fox and Stryker, 2017; Pezzulo et al., 2015; Toyoizumi et al., 2014; Turrigiano, 2017). To what extent seemingly indifferent and/or inconsistent choices may eventually be explained away with these and/or similar biological constraints is an open and challenging issue.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. NeuroImage *31*, 790–795.

Aitchison, L., and Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. Curr. Opin. Neurobiol. *46*, 219–227.

Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nat. Rev. Neurosci. *7*, 358–366.

Barlow, H. (1961). Possible Principles Underlying the Transformations of Sensory Messages. Sens. Commun. *1*.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical Microcircuits for Predictive Coding. Neuron *76*, 695–711.

Borst, J.P., Taatgen, N.A., and van Rijn, H. (2011). Using a symbolic process model as input for model-based fMRI analysis: Locating the neural correlates of problem state replacements. NeuroImage *58*, 137–147.

Botvinik-Nezer, R., Iwanir, R., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Dreber, A., Camerer, C.F., Poldrack, R.A., and Schonberg, T. (2019). fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study. Sci. Data *6*, 106.

Botvinik-nezer, R., Holzmeister, F., Camerer, C.F., and Johannesson, M. (2019). Variability in the analysis of a single neuroimaging dataset by many teams.

Brenner, N., Bialek, W., and Steveninck, R. de R. van (2000). Adaptive Rescaling Maximizes Information Transmission. Neuron *26*, 695–702.

Burke, C.J., Baddeley, M., Tobler, P.N., and Schultz, W. (2016). Partial Adaptation of Obtained and Observed Value Signals Preserves Information about Gains and Losses. J. Neurosci. Off. J. Soc. Neurosci. *36*, 10016–10025.

Buschman, T.J., Siegel, M., Roy, J.E., and Miller, E.K. (2011). Neural substrates of cognitive capacity limitations. Proc. Natl. Acad. Sci. U. S. A. *108*, 11252–11255.

Christopoulos, G.I., Tobler, P.N., Bossaerts, P., Dolan, R.J., and Schultz, W. (2009). Neural Correlates of Value, Risk, and Risk Aversion Contributing to Decision Making under Risk. J. Neurosci. *29*, 12574–12583.

Conen, K.E., and Padoa-Schioppa, C. (2019). Partial Adaptation to the Value Range in the Macaque Orbitofrontal Cortex. J. Neurosci. *39*, 3498–3513.

Cox, K.M., and Kable, J.W. (2014). BOLD subjective value signals exhibit robust range adaptation. J. Neurosci. Off. J. Soc. Neurosci. *34*, 16533–16543.

Dale, A.M. (1999). Optimal experimental design for event-related fMRI. Hum. Brain Mapp. *8*, 109–114.

Daunizeau, J. (2017). The variational Laplace approach to approximate Bayesian inference. ArXiv170302089 Q-Bio Stat.

Daunizeau, J. (2019). Variational Bayesian modelling of mixed-effects. ArXiv190309003 Cs Stat.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLoS Comput Biol *10*, e1003441.

Dayan, P., and Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. *8*, 429–453.

Deco, G., Rolls, E.T., Albantakis, L., and Romo, R. (2013). Brain mechanisms for perceptual and reward-related decision-making. Prog. Neurobiol. *103*, 194–213.

Diederen, K.M.J., Spencer, T., Vestergaard, M.D., Fletcher, P.C., and Schultz, W. (2016). Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency. Neuron *90*, 1127–1138.

Diedrichsen, J., and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLOS Comput. Biol. *13*, e1005508.

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H.H., and Kriegeskorte, N. (2020). Comparing representational geometries using the unbiased distance correlation. ArXiv200702789 Stat.

Dockès, J., Poldrack, R., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., Thirion, B., and Varoquaux, G. (2020). NeuroQuery: comprehensive meta-analysis of human brain mapping. 1–34.

Doya, K., Ishii, S., Pouget, A., and Rao, R.P.N. (2007). Bayesian Brain: Probabilistic Approaches to Neural Coding (MIT Press).

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. Neuron *92*, 1398–1411.

Elliott, R., Agnew, Z., and Deakin, J.F.W. (2008). Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. Eur. J. Neurosci. *27*, 2213–2218.

Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898–1902.

Fox, K., and Stryker, M. (2017). Integrating Hebbian and homeostatic plasticity: introduction. Philos. Trans. R. Soc. B Biol. Sci. *372*.

Frank, M.J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw. Off. J. Int. Neural Netw. Soc. *19*, 1120–1136.

Friston, K. (2012). The history of the future of the Bayesian brain. NeuroImage *62*, 1230–1233.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the Laplace approximation. NeuroImage *34*, 220–234.

Friston, K.J., Diedrichsen, J., Holmes, E., and Zeidman, P. (2019). Variational representational similarity analysis. NeuroImage *201*, 115986.

Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. *37*, 1297–1310.

Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1986). Neuronal population coding of movement direction. Science *233*, 1416–1419.

Güçlü, U., and Gerven, M.A.J. van (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J. Neurosci. *35*, 10005–10014.

Hebb, D.O. (1950). A review of "The organization of behaviour: A neuropsychological theory." Q. J. Exp. Psychol. *2*, 142–143.

Henson, R. (2006). Forward inference using functional neuroimaging: dissociations versus associations. Trends Cogn. Sci. *10*, 64–69.

Hopfinger, J.B., Büchel, C., Holmes, A.P., and Friston, K.J. (2000). A Study of Analysis Parameters That Influence the Sensitivity of Event-Related fMRI Analyses. NeuroImage *11*, 326–333.

Hosoya, T., Baccus, S.A., and Meister, M. (2005). Dynamic predictive coding by the retina. Nature *436*, 71–77.

Jocham, G., Hunt, L.T., Near, J., and Behrens, T.E.J. (2012). A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. Nat. Neurosci. *15*, 960–961.

Kietzmann, T.C., McClure, P., and Kriegeskorte, N. (2017). Deep Neural Networks in Computational Neuroscience. BioRxiv 133504.

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. *116*, 21854–21863.

Kobayashi, S., Pinto de Carvalho, O., and Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. J. Neurosci. Off. J. Soc. Neurosci. *30*, 534–544.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci.

Kriegeskorte, N., and Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *371*.

Kriegeskorte, N., and Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. Annu. Rev. Neurosci. *42*, 407–432.

Kriegeskorte, N., and Golan, T. (2019). Neural network models and deep learning. Curr. Biol. *29*, R231–R236.

Kuhnen, C.M., and Knutson, B. (2005). The neural basis of financial risk taking. Neuron *47*, 763–770.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. Z. Naturforsch. [C] *36*, 910–912.

Lewicki, M.S. (2002). Efficient coding of natural sounds. Nat. Neurosci. *5*, 356–363.

Lisman, J. (2017). Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes: long-term potentiation, long-term depression, short-term potentiation and scaling. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *372*.

Louie, K., and Glimcher, P.W. (2012). Efficient coding and the neural representation of value. Ann. N. Y. Acad. Sci. *1251*, 13–32.

Marois, R., and Ivanoff, J. (2005). Capacity limits of information processing in the brain. Trends Cogn. Sci. *9*, 296–305.

Martens, M.B., Celikel, T., and Tiesinga, P.H.E. (2015). A Developmental Switch for Hebbian Plasticity. PLOS Comput. Biol. *11*, e1004386.

Martino, B.D., Kumaran, D., Seymour, B., and Dolan, R.J. (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. Science *313*, 684–687.

Martino, B.D., Camerer, C.F., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. Proc. Natl. Acad. Sci. *107*, 3788–3792.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. Minds Mach. *1*, 185–196.

Miller, E.K., and Buschman, T.J. (2015). Working Memory Capacity: Limits on the Bandwidth of Cognition. Daedalus *144*, 112–122.

O'Doherty, J.P., Hampton, A., and Kim, H. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. Ann. N. Y. Acad. Sci. *1104*, 35–53.

Padoa-Schioppa, C. (2009). Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. J. Neurosci. *29*, 14004–14014.

Penny, W.D., and Ridgway, G.R. (2013). Efficient Posterior Probability Mapping Using Savage-Dickey Ratios. PLOS ONE *8*, e59655.

Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., and Leff, A.P. (2010). Comparing Families of Dynamic Causal Models. PLoS Comput Biol *6*, e1000709.

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. Prog. Neurobiol. *134*, 17–35.

Polanía, R., Woodford, M., and Ruff, C.C. (2019). Efficient coding of subjective value. Nat. Neurosci. *22*, 134–142.

Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., and Milham, M.P. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. Front. Neuroinformatics *7*, 12.

Ramsey, N.F., Jansma, J.M., Jager, G., Van Raalten, T., and Kahn, R.S. (2004). Neurophysiological factors in human information processing capacity. Brain J. Neurol. *127*, 517–525.

Rangel, A., and Clithero, J.A. (2012). Value normalization in decision making: theory and evidence. Curr. Opin. Neurobiol. *22*, 970–981.

Rigoux, L., and Daunizeau, J. (2015). Dynamic causal modelling of brain-behaviour relationships. NeuroImage *117*, 202–221.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C.D., Tobler, P.N., and Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. Philos. Trans. R. Soc. B Biol. Sci. *363*, 3801–3811.

Sharot, T. (2011). The optimism bias. Curr. Biol. *21*, R941–R945.

Sharot, T., Korn, C.W., and Dolan, R.J. (2011). How unrealistic optimism is maintained in the face of reality. Nat. Neurosci. *14*, 1475–1479.

Shouval, H.Z., Wang, S.S.-H., and Wittenberg, G.M. (2010). Spike Timing Dependent Plasticity: A Consequence of More Fundamental Learning Rules. Front. Comput. Neurosci. *4*.

Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annu. Rev. Neurosci. *24*, 1193–1216.

Soltani, A., Martino, B.D., and Camerer, C. (2012). A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence. PLOS Comput. Biol. *8*, e1002607.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge, Mass.: A Bradford Book).

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. Science *315*, 515–518.

Toyoizumi, T., Kaneko, M., Stryker, M.P., and Miller, K.D. (2014). Modeling the Dynamic Interaction of Hebbian and Homeostatic Plasticity. Neuron *84*, 497–510.

Turner, B.M., Forstmann, B.U., Wagenmakers, E.-J., Brown, S.D., Sederberg, P.B., and Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. NeuroImage *72*, 193–206.

Turner, B.M., Rodriguez, C.A., Norcia, T.M., McClure, S.M., and Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. NeuroImage *128*, 96–115.

Turner, B.M., Palestro, J.J., Miletić, S., and Forstmann, B.U. (2019). Advances in techniques for imposing reciprocity in brain-behavior relations. Neurosci. Biobehav. Rev. *102*, 327–336.

Turrigiano, G.G. (2017). The dialectic of Hebb and homeostasis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *372*.

Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. Neuron *60*, 215–234.

Wark, B., Lundstrom, B.N., and Fairhall, A. (2007). Sensory adaptation. Curr. Opin. Neurobiol. *17*, 423–429.

Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference (New York: Springer-Verlag).

Wei, X.-X., and Stocker, A.A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. Nat. Neurosci. *18*, 1509–1517.

Wyart, V., and Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. Curr. Opin. Behav. Sci. *11*, 109–115.

Zenke, F., and Gerstner, W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. Philos. Trans. R. Soc. B Biol. Sci. *372*, 20160259.

Zimmermann, J., Glimcher, P.W., and Louie, K. (2018). Multiple timescales of normalized value coding underlie adaptive choice behavior. Nat. Commun. *9*, 3206.

Appendix 1: range adaptation

In what follows we provide the mathematical derivation of equation (6) of the main text.

Let σ^* be the value of the pseudo-variance parameter σ that maximizes the derivative of a given ANN unit activation function w.r.t. to its inputs, i.e.:

$$\sigma^* = \arg\max_{\sigma} \left| \frac{\partial f_.}{\partial u} \Big|_{u,\sigma} \right| \tag{A1}$$

where $f_.$ is the activation function of neural units in the ANN (cf. Equation 2), and we have dropped unit indices for mathematical convenience.

Range adaptation proceeds by modifying the pseudo-variance parameter in the direction of σ^*, with a step size that is controlled by the learning rate α_{RA}, i.e.:

$$\sigma_{t+1} = \sigma_t + \alpha_{RA} \times \left(\sigma^* - \sigma_t \right) \tag{A2}$$

Setting $\alpha_{RA} < 1$ ensures that the pseudo-variance parameter integrates the history of past inputs when adapting its range of activation.

Let us first focus on pseudo-gaussian activation functions. Without loss of generality, we will drop the time index and use a centred input $\tilde{u} = u - \mu$. The first derivative of the activation function is given by:

$$\frac{\partial f_{Gauss}}{\partial u}\Big|_{\tilde{u},\mu,\sigma} = \frac{2\tilde{u}}{\sigma^2} \exp\left(-\frac{\tilde{u}}{\sigma^2} \right) \tag{A3}$$

Range adaptation proceeds by maximizing Equation A3 with respect to σ, which reduces to finding the zero of the mixed partial derivative of f_{Gauss}:

$$\left.\frac{\partial^2 f_{Gauss}}{\partial\sigma\,\partial\tilde{u}}\right|_{\tilde{u},\mu,\sigma} = \left(\sigma^2 - \tilde{u}^2\right)\frac{4\tilde{u}}{\sigma^5}\exp\left(-\frac{\tilde{u}^2}{\sigma^2}\right) \tag{A4}$$

Since σ is positive, the maximum of Equation A3 is simply given by $\sigma^* = \left|u\right| = \left|u-\mu\right|$. Inserting the expression for σ^* into Equation A2 then provides the following learning rule:

$$\sigma_{t+1} = \sigma_t + \alpha_{RA}\times\left(\left|\mu - u_t\right| - \sigma_t\right) \tag{A5}$$

Noting the input to the integration layer is given by $\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} C^{(i,j,k)}x_t^{(i,j)}$ eventually yields Equation 6 of the main text.

Let us now focus on sigmoid activation functions. In this case, we use the following change of variable: $u = \left(u-\mu\right)*\gamma$, where γ is the arbitrary scaling factor of the sigmoid activation function (cf. Equation 2). We will see that it is possible to set such that the range adaptation learning rule is identical for both pseudo-gaussian and sigmoid activation functions.

It is trivial to show that the first and mixed partial derivative of the sigmoid activation function are given by:

$$\begin{aligned}\left.\frac{\partial f_{sigmoid}}{\partial\tilde{u}}\right|_{\tilde{u},\mu,\sigma} &= \frac{1}{\sigma}\frac{\mathrm{e}^{-\tilde{u}/\sigma}}{\left(\mathrm{e}^{-\tilde{u}/\sigma}+1\right)^2}\\ \left.\frac{\partial^2 f_{sigmoid}}{\partial\sigma\,\partial\tilde{u}}\right|_{\tilde{u},\mu,\sigma} &= \mathrm{e}^{-\tilde{u}/\sigma}\frac{-\left(\tilde{u}/\sigma+1\right)\mathrm{e}^{-\tilde{u}/\sigma}+\left(\tilde{u}/\sigma-1\right)}{\sigma^2\left(\mathrm{e}^{-\tilde{u}/\sigma}+1\right)^3}\end{aligned} \tag{A6}$$

Finding the zero of the mixed partial derivative reduces to solving $(y-1)-e^{-y}(y+1)$ with $y = \tilde{u}/\sigma$. There is no analytical closed form solution to this equation, but a numerical approach yields $y^* \approx \pm 1.5434$. Since σ is positive, the solution is simply given by $\sigma^* = \left|\tilde{u}/y^*\right| = \left|u-\mu\right|\times\gamma/1.5434$. Setting $\gamma = 1.5434$ then simplifies the solution to $\sigma^* = \left|u-\mu\right|$, which thus provides the same learning rule as Equation A3 above.

Appendix 2: fMRI results statistics

In what follows, we provide summary statistics of our ANN-based behavioural and fMRI analyses. Table 1 gives the mean percentage of explained behavioral variance (R^2) and it standard deviation (across participants) for each model, for both groups. Table 2 gives the mean R^2 difference between each ANN model and the reference logistic model, its standard deviation, and the resulting p-value (H_0: no R^2 difference, dof=53), for both groups. Table 3 gives the p-value of RDM correlations for each model and each ROI (H_0: $\rho \leq 0$, one-sided t-test, dof=53), in the 'equal' range' group. Table 4 gives the p-value of RDM correlations for each model and each ROI (H_0: $\rho \leq 0$, one-sided t-test, dof=53), in the 'equal indifference' group. In all tables, statistical significance is highlighted in green (with the appropriate threshold correction for multiple comparisons).

	'equal range'		'equal indifference'	
	mean	std	mean	std
logistic	0.7789	0.1049	0.7367	0.1273
G-ANN	0.8451	0.0784	0.8831	0.0866
S-ANN	0.8374	0.0793	0.7845	0.1166
G-RA-ANN	0.8636	0.0682	0.7937	0.1822
S-RA-ANN	0.8402	0.0878	0.8409	0.1094
G-H-ANN	0.8621	0.085	0.867	0.1673
S-H-ANN	0.8527	0.0623	0.8225	0.127

Table 1: Mean R^2 and its standard deviation for each model, for both groups.

	'equal range'			'equal indifference'		
	mean	std	p-value	mean	std	p-value
G-ANN	0.0662	0.0939	5.00E-06	0.1463	0.0748	3.00E-20
S-ANN	0.0586	0.0446	1.00E-12	0.0477	0.0311	6.00E-16
G-RA-ANN	0.0847	0.0725	6.00E-11	0.0569	0.1884	1.53E-02
S-RA-ANN	0.0614	0.0363	4.00E-16	0.1041	0.0669	4.00E-16
G-H-ANN	0.0832	0.0601	3.00E-13	0.1302	0.167	3.00E-07
S-H-ANN	0.0738	0.0841	8.00E-08	0.0857	0.119	2.00E-06

Table 2: Mean R^2 difference and its standard deviation for each ANN model, for both groups.

	G-ANN	S-ANN	G-RA-ANN	S-RA-ANN	G-H-ANN	S-H-ANN
Motor_L	0.9985	0.9684	0.0056	0.9532	0.0096	0.0455
Motor_M	0.9909	0.998	0.1474	0.911	0.0022	0.0008
Motor_R	0.9994	0.9558	0.0335	0.7906	0.0113	0.0195
Visual_L	0.9978	0.9768	0.01	0.9658	0.0684	0.2107
Visual_M	0.9146	0.7227	0.0762	0.9782	0.184	0.5914
Visual_R	0.9458	0.8604	0.0707	0.8023	0.0226	0.0626
PCC	0.9737	0.9226	0.0644	0.9616	0.0101	0.1186
dACC	0.9999	0.9868	0.1332	0.7359	0.0017	0.0008
ACC	0.9996	0.9487	0.2128	0.9518	0.0014	0.0037
dLPFC_L	0.9856	0.7515	0.0271	0.9953	0.004	0.0082
dLPFC_R	0.9887	0.7984	0.0034	0.7863	0.0004	0.0004
vmPFC	0.9999	0.8739	0.1146	0.9821	0.0407	0.0535
Insula_L	0.9965	0.9485	0.0507	0.9592	0.0004	3.57E-05
Insula_R	0.9642	0.9376	0.1517	0.9201	0.0052	0.0067
Amygdala_L	0.9889	0.5912	0.4413	0.983	0.0006	0.0001
Amygdala_R	0.9773	0.9464	0.2797	0.9175	1.51E-05	9.45E-07
Striatum_L	0.9986	0.716	0.3392	0.8637	1.86E-05	3.13E-06
Striatum_R	0.9956	0.9768	0.222	0.9845	0.0001	0.0002

Table 3: P-value of RDM correlations for each model and each ROI ('equal' range').

	G-ANN	S-ANN	G-RA-ANN	S-RA-ANN	G-H-ANN	S-H-ANN
Motor_L	0.4258	0.9914	0.0784	0.7568	0.1149	1.88E-06
Motor_M	0.5736	0.9907	0.0324	0.839	0.0066	1.94E-06
Motor_R	0.6155	0.9975	0.0559	0.6253	0.0336	7.20E-07
Visual_L	0.6878	0.9954	0.153	0.5337	0.2292	2.41E-05
Visual_M	0.7691	0.9556	0.0014	0.0344	0.9672	3.22E-03
Visual_R	0.5845	0.9997	0.0249	0.5964	0.0202	1.22E-06
PCC	0.2997	0.9995	0.1245	0.6836	0.4924	9.47E-05
dACC	0.455	0.9998	0.1214	0.8494	0.0097	1.00E-07
ACC	0.0753	0.9993	0.1022	0.8217	0.0026	6.30E-07
dLPFC_L	0.051	0.9995	0.0146	0.4022	0.4083	1.94E-04
dLPFC_R	0.3554	0.9998	0.0249	0.9613	0.0037	8.10E-07
vmPFC	0.3914	0.9742	0.0148	0.7574	0.1307	6.89E-05
Insula_L	0.2355	0.9684	0.3209	0.964	0.0009	1.00E-08
Insula_R	0.3882	0.9999	0.0135	0.8121	0.0171	3.71E-05
Amygdala_L	0.4186	0.9998	0.3852	0.9786	0.0026	2.00E-08
Amygdala_R	0.7117	0.9987	0.0353	0.9971	2.00E-04	3.00E-08
Striatum_L	0.5835	0.9997	0.1023	0.9974	2.00E-04	4.00E-08
Striatum_R	0.3378	0.9998	0.0955	0.9782	0.0001	3.30E-07

Table 4: P-value of RDM correlations for each model and each ROI ('equal' indifference).