Irrational decisions reflect robustness constraints on value computations implemented by orbitofrontal circuits

Juliette Bénon[1], Mathias Pessiglione[1], Fabien Vinckier[1], Jean Daunizeau[1]

[1] Paris Brain Institute, Paris, France

Abstract

We first train cohorts of artificial neural nets to perform ten variants of **rational** decision-relevant computations. Those variants that operate under a specific option-encoding format exhibit most of the electrophysiological coding properties observed in orbitofrontal neurons of monkeys making decisions under risk **1**.

We then distort these neural nets' internal wiring to reproduce monkeys' **irrational** choices. This induces deterministic spillover interferences in decision-relevant computations **2** that generalize across individuals, at both the behavioral and neural level **3**.

Importantly, although irrational nets do not seem to bring informational or metabolic benefits, they display enhanced tolerance to damage and noise when compared to their rational counterparts **4**.

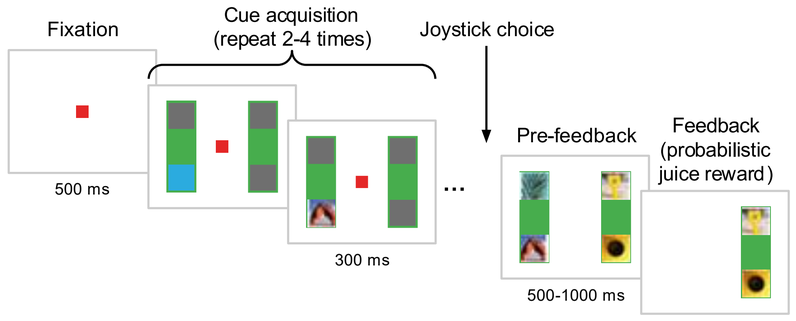
➔ This suggests that **some forms of irrational behavior may be the incidental outcome of distal evolutionary pressure on the robustness of orbitofrontal circuits**.

Methods

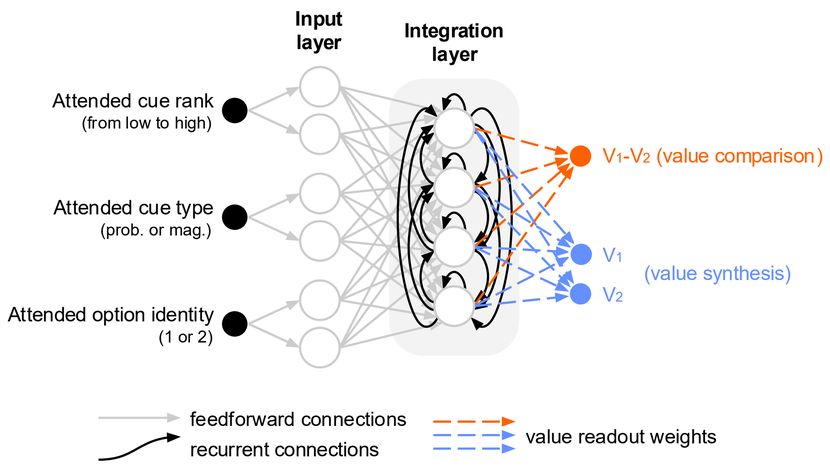
Choice task between probabilistic rewards

Dataset (task, behaviour, electrophysiological records) collected by **Hunt et al., 2018**.

Hunt, L. T., Malalasekera, W. M. N., de Berker, A. O., Miranda, B., Farmer, S. F., Behrens, T. E. J., & Kennerley, S. W. (2018). Triple dissociation of attention and decision computations across prefrontal cortex. *Nature Neuroscience*, *21*(10), 1471–1481

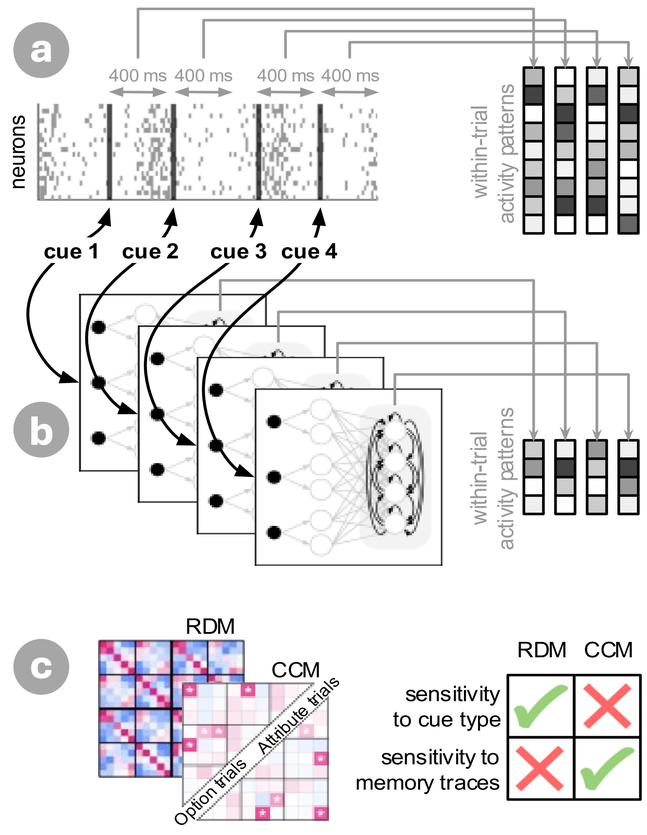


Recurrent neural networks (RNNs)



Comparing RNNs to neural data

Electrophysiological recordings **(a)** and internal RNN activity **(b)** are compared using two measures derived from separate representational geometry matrices **(c)**.



1 Identifying idealised models of the orbitofrontal cortex (OFC)

Two models reproduce OFC neurons' electrophysiological activity better than expected by chance:

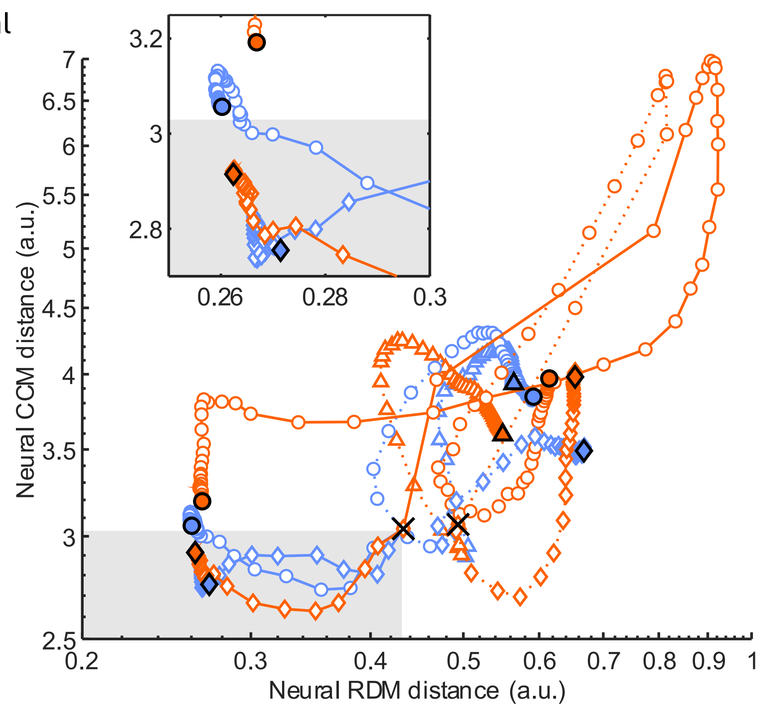
Comparison model

Cue about the **first** or **second** option → Value of the **attended** option − Value of the **unattended** option

Synthesis model

Cue about the **first** or **second** option → Value of the **attended** option & Value of the **unattended** option

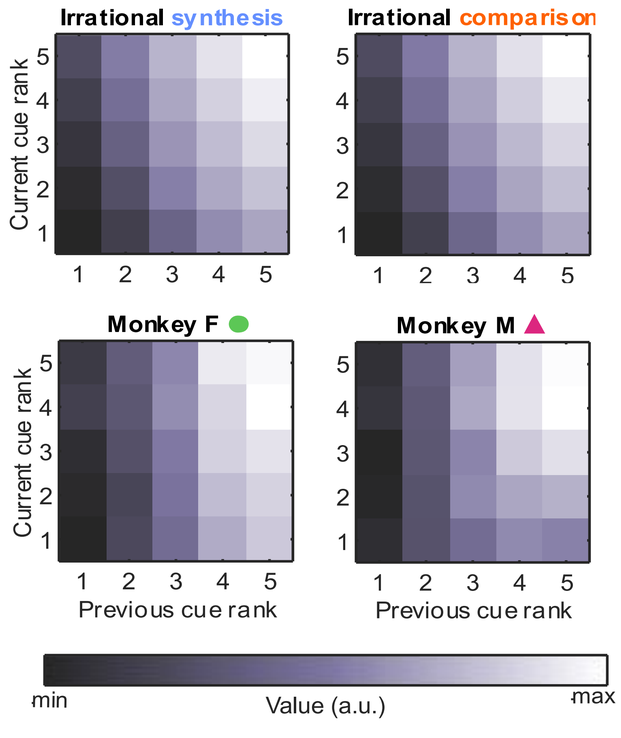
network value computation: synthesis, comparison
attended option identity format: spatial, temporal
value readout format: spatial, temporal, attentional



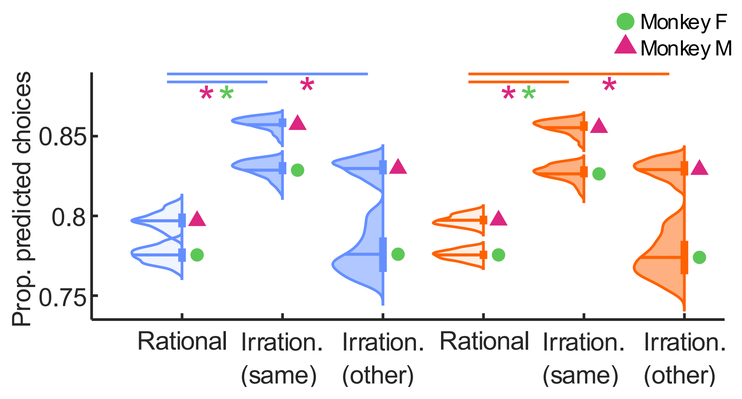
2 Deterministic interferences produce irrational choices

Irrational RNNs and monkeys weigh the previous cue more than the current one when estimating the current option's value.

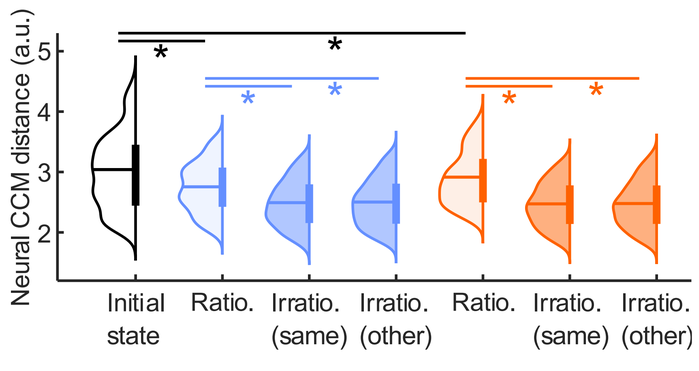
Interference effects accumulate within trials.



3 Irrational RNNs are both behaviourally and neurally realistic



Behavioural predictions of irrational RNNs outperform rational RNNs, and generalize across trials and individuals.



Irrational RNNs better resemble OFC activity than rational RNNs. This improvement generalizes across individuals.

4 Irrational models of the OFC are more resilient to damage and noise than their rational counterparts

Irrational RNNs (♦, ▲) retain more rational choices than their rational counterparts (●) under:

(a) Lesions of integration layer units
(b) Disconnections of recurrent connections
(c) Neural noise added to integration layer units