The Roles of Online and Offline Replay in Planning

Eldar *et al*., 2020

Presented by Gabriela Iwama & Lena Mehnert

Agenda

Introduction

- Background on replay
- Aim of study

Methods

- Experimental task and MEG

——— **BREAK** ———

Results

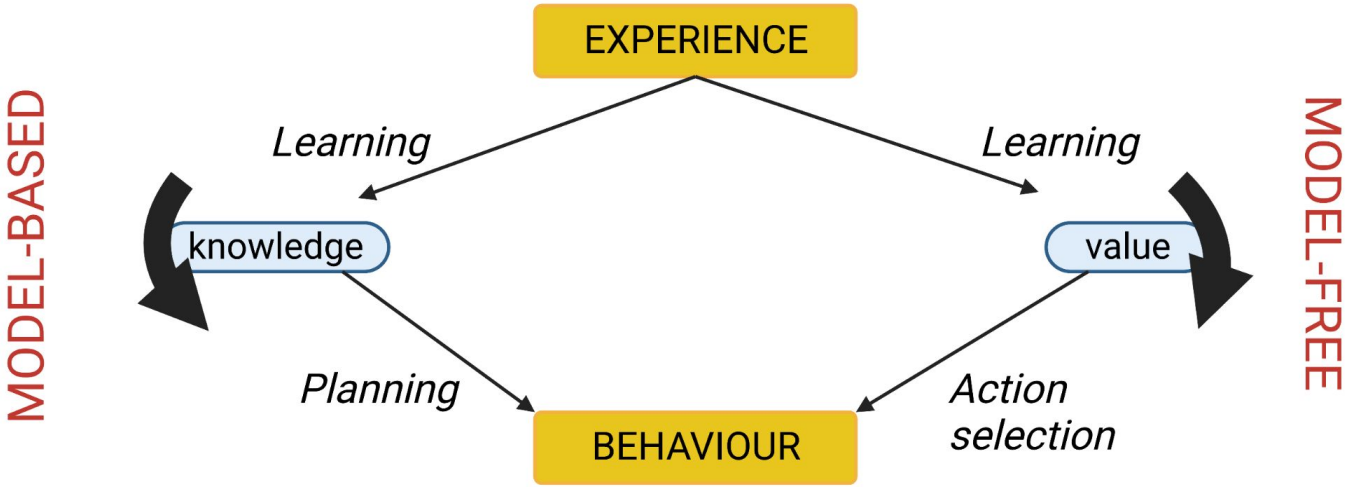
- Role of replay in planning

Discussion of the paper

- Summary and conclusion
- Limitations

Open discussion with you

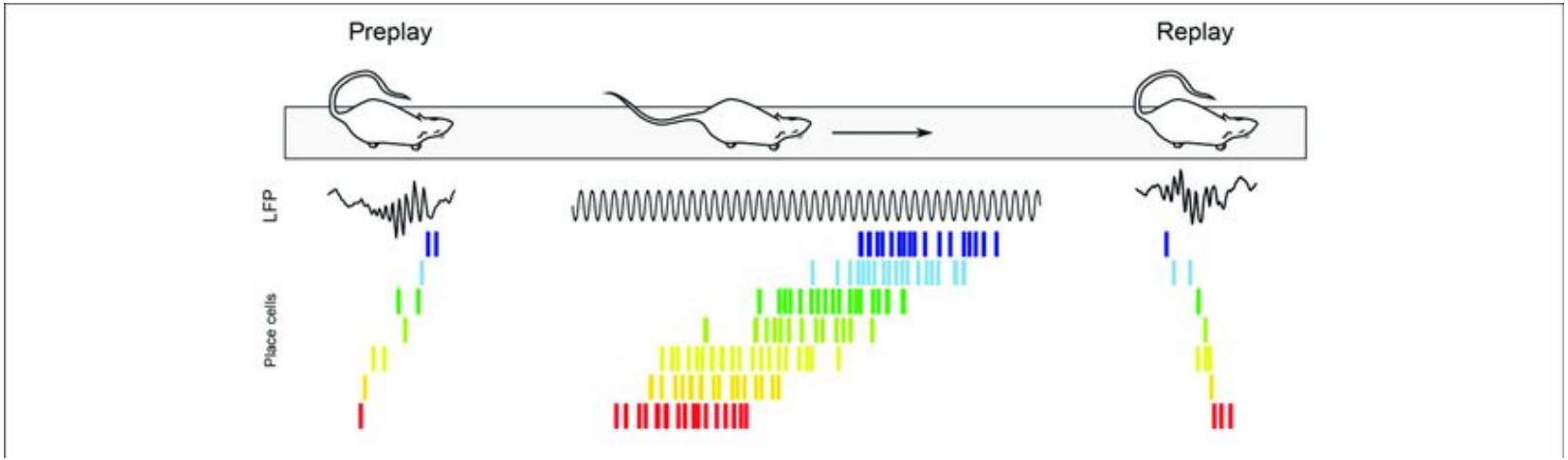
Overview



Adapted from: Bera *et al.*, 2019
Created with BioRender.com

INTRO **What is replay?**

Replay: a pattern of electrical activity is played back in order

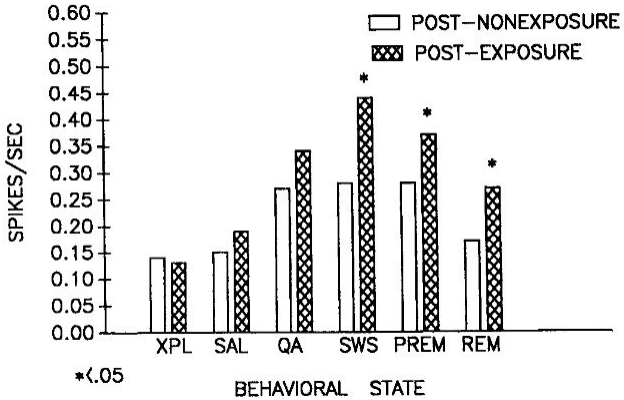


From: Drieu and Zugaro, 2019

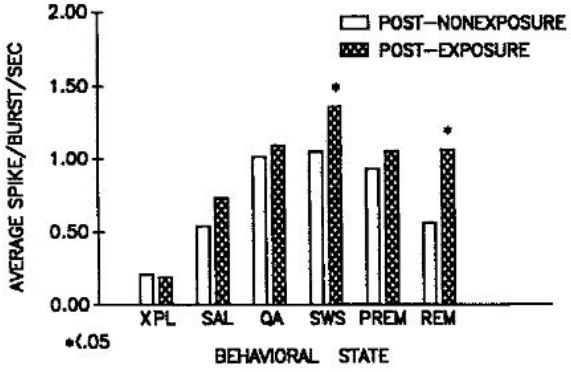
INTRO **Origins of replay**

Hippocampus: important for memory and spatial navigation
Replay: internally generated neuronal sequences in the hippocampus (place cells)

Pavlides and Winson, 1989



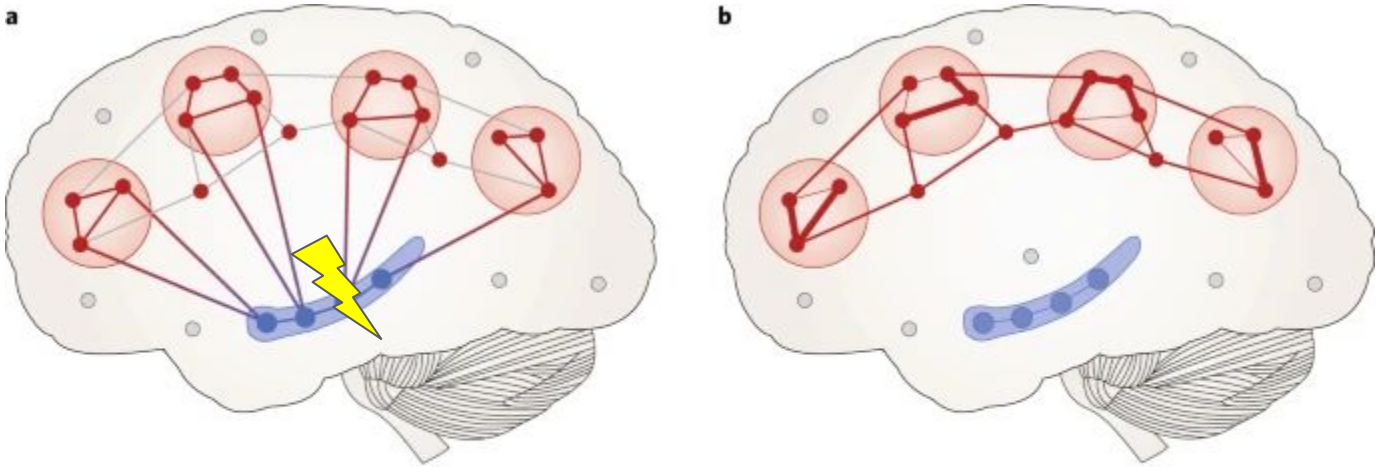
significantly increased firing rate of post-exposure CA1 neurons in sleep states



higher number of spike bursts during SWS and REM

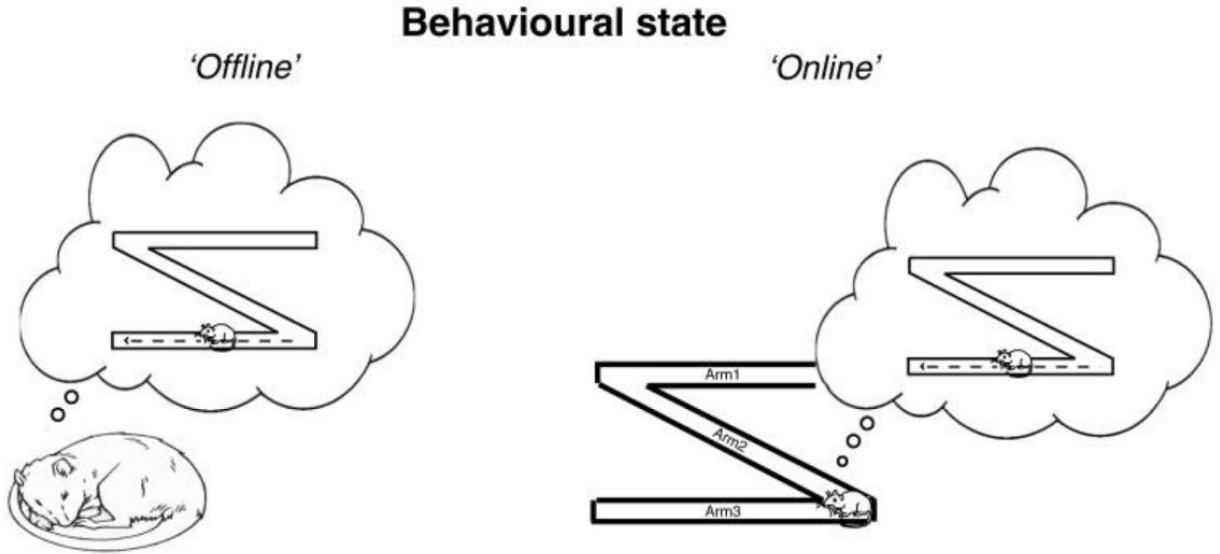
INTRO **Replay and Memory**

MEMORY CONSOLIDATION



→ REPLAY in spatial memory consolidation and memory retrieval
→ bilateral lesions in hippocampus leads to strong impairment of spatial memory

INTRO **OFF vs. ON replay**



after an experience has happened
→ during rest or sleep

while the experience is still happening
→ during a task (awake state)

INTRO **Replay in planning**

ONLINE REPLAY
Model-based policy
→ evaluation based on reward states
→ on-the-fly planning

OFFLINE REPLAY
(1) complementary role
→ specifying routes and rewards
→ flexible decision making
(2) model-free decision policy
→ less flexible
→ pre-formulated

SUGGESTION
Trade-off between ON and OFF replay
→ ON: model-based flexibility
→ OFF: model-free policy

HYPOTHESIS **Aim of study**

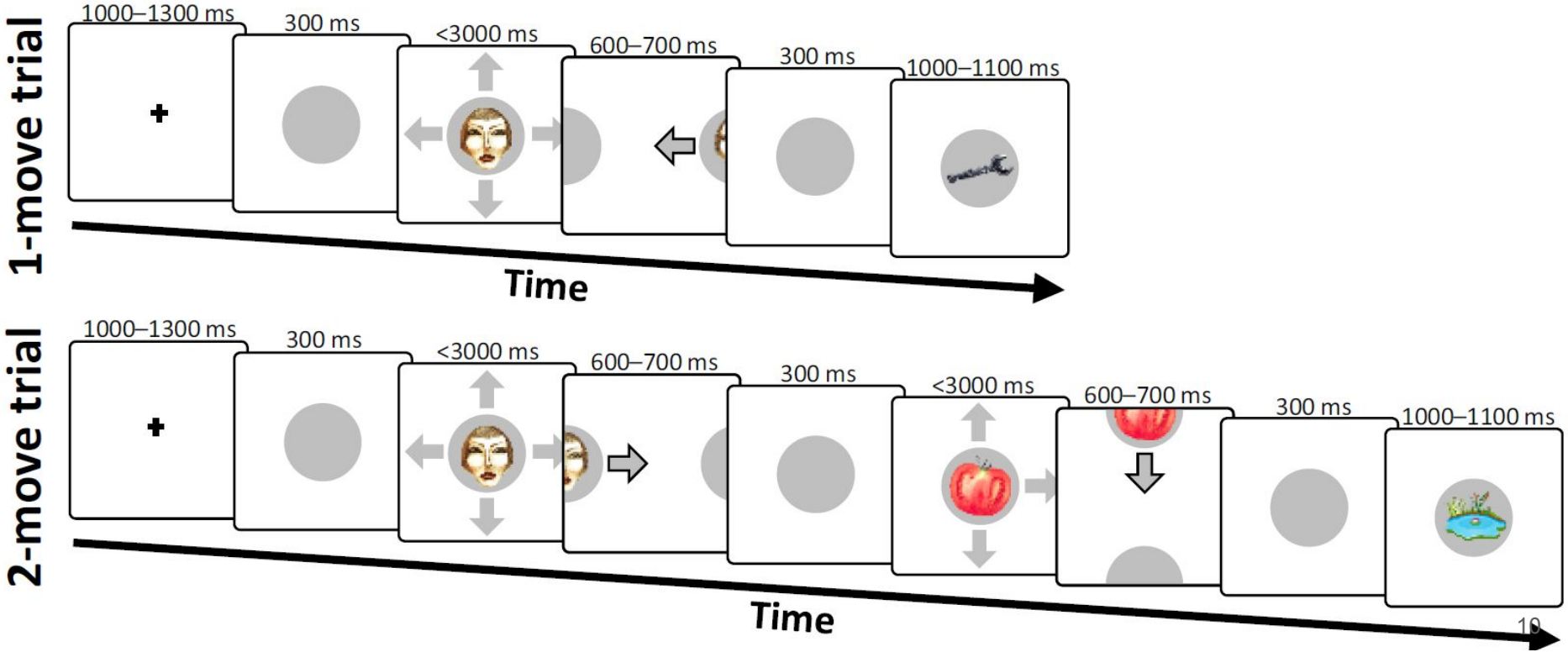
How do online and offline replay contribute to planning and decision making?

What are the neural correlates?

Do online and offline replay have complementary or contrasting impacts?

What is the correlation to behaviour and behavioural disorders?

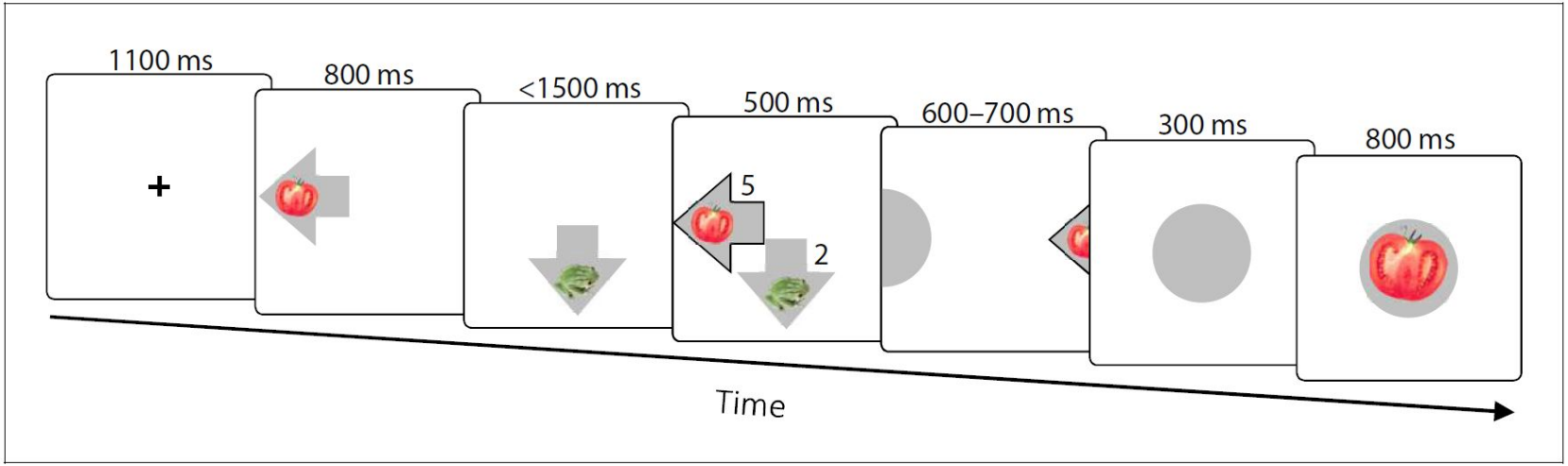
METHOD **Behavioral Task**



METHOD State Space

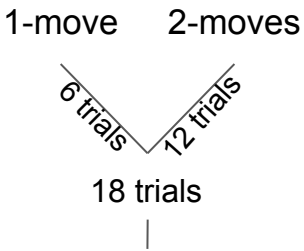
METHOD **Training Phase**

State-Reward Training



State-Space Training

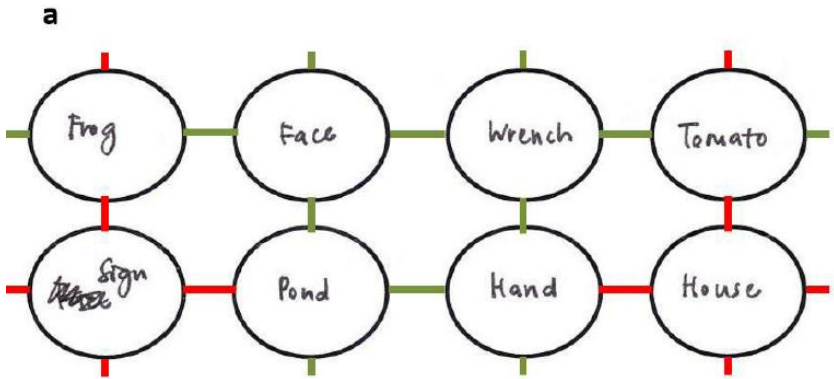
METHOD Exp. Design

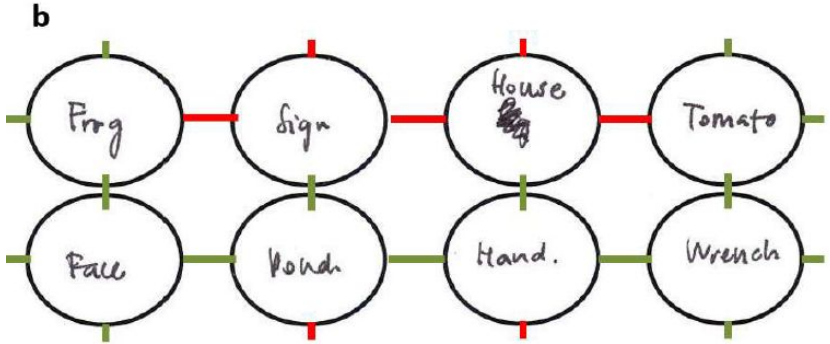


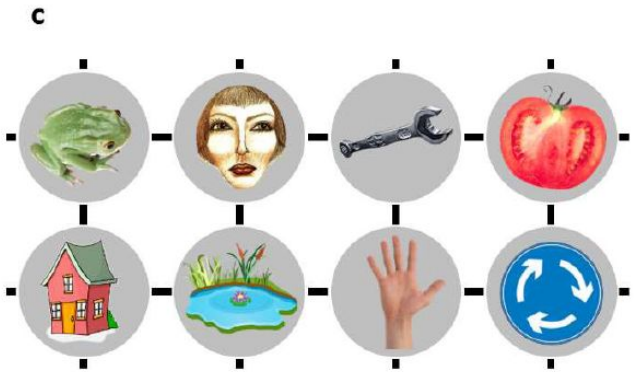
Training R F F F R N F F **Reward** R N F F R N F F **Spatial** R N F F

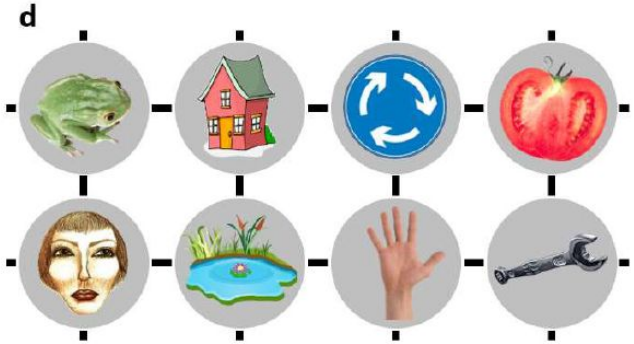
1 2 3 4 5

METHOD **Exp. Design**



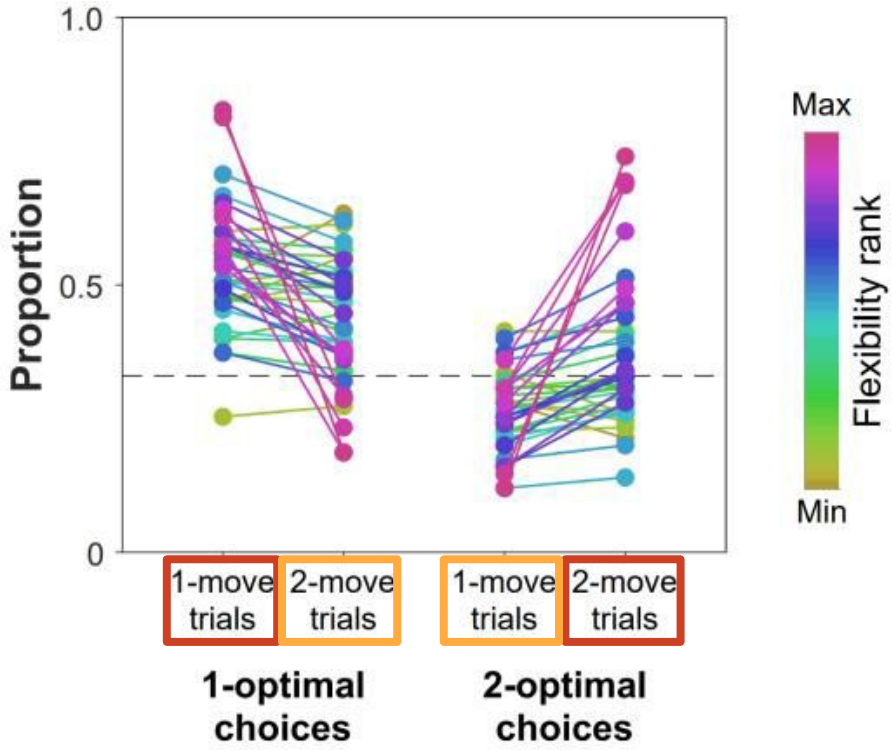






METHOD Individual Flexibility

$$IF = \%\ optimal\ choices\ in\ current\ cond - \%\ optimal\ choices\ in\ other\ cond$$

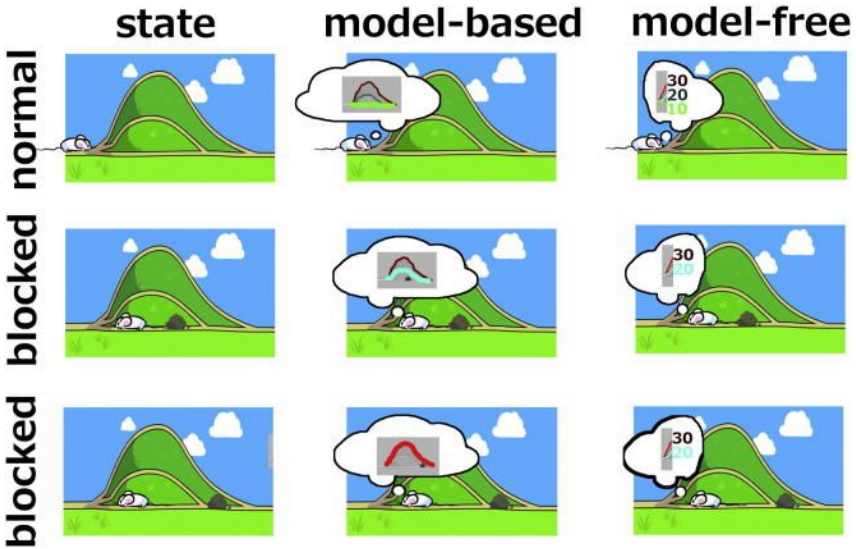


METHOD **Learning Algorithms**

Model-free

Model-based

Hybrid MF-MB



Dolan, Dayan (2013)

METHOD **Learning Algorithms**

Model-free algorithm

- Q^{MF1} and Q^{MF2} - Q-values
- η^{MF1} and η^{MF1} - learning rate
- τ - memory parameter - control decay to theta
- τ' - memory parameter after reward/spatial change
- γ - Bias parameter towards a move

$$Q_{t+1}^{\mathrm{MF1}}\left(s_{t,1},m_t\right)=Q_t^{\mathrm{MF1}}\left(s_{t,1},m_t\right)+\eta^{\mathrm{MF1}}\delta_t^{\mathrm{MF1}}$$

$$Q^{\mathrm{MF}}\leftarrow\tau^{\mathrm{MF}}Q^{\mathrm{MF}}+\left(1-\tau^{\mathrm{MF}}\right)\theta$$

$$\mathrm{p}\left(m_t=a|s_t\right)\propto e^{\gamma_m+\beta_1^{\mathrm{MF1}}Q_t^{\mathrm{MF1}}\left(s_{t,1},m\right)}$$

$$\mathrm{p}\left(m_{t,1}=m|s_{t,1}\right)\propto e^{\gamma_m+\beta_2^{\mathrm{MF1}}Q_t^{\mathrm{MF1}}\left(s_{t,1},m\right)+\beta_2^{\mathrm{MF2}}Q_t^{\mathrm{MF2}}\left(s_{t,1},m\right)}$$

METHOD Learning Algorithms

Model-based algorithm

- T - state transitions
- **ρ** - infer opposite transition
- τ - memory parameter - control decay to chance level
- τ' - memory after reward/spatial change
- Q^{MB} - integrating the probability and reward of all potential images
- κ - weight for second move reward

$$T_{t+1}\left(s_{t,1}, m_t, s_{t,2}\right) = T_t\left(s_{t,1}, m_t, s_{t,2}\right) + \eta^{\mathrm{MB}}\delta_t^{\mathrm{MB}}$$

$$T_{t+1}\left(s_{t,2}, \tilde{m}_t, s_{t,1}\right) = T_t\left(s_{t,2}, \tilde{m}_t, s_{t,1}\right) + \rho\eta^{\mathrm{MB}}\delta_t'^{\mathrm{MB}}$$

$$T \leftarrow \tau^{\mathrm{MB}} T + \left(1 - \tau^{\mathrm{MB}}\right)\frac{1}{7}$$

For 1-move trials:

$$Q_t^{MB}\left(s_{t,1}, m\right) = \sum_s T_t\left(s_{t,1}, m, s\right) R_g(s)$$

For 2-move trials:

$$Q_t^{MB}\left(s_{t,1}, m\right) = \sum_s T_t\left(s_{t,1}, m, s\right)\left(R_g(s) + \kappa \max_{m'} \sum_{s'} T_t\left(s, m', s'\right) R_g\left(s'\right)\right)$$

$$\mathrm{p}\left(m_t = m \middle| s_{t,1}\right) \propto e^{\gamma_m + \beta^{\mathrm{MB}} Q_t^{\mathrm{MB}}\left(s_{t,1}, m\right)}$$

METHOD Learning Algorithms

Hybrid MB-MF algorithm

1-move trials:

$$\mathrm{p}(m_t = m|s_t) \propto e^{\gamma_m + \beta_1^{\mathrm{MF1}} Q_t^{\mathrm{MF1}}(s_{t,1},m) + \beta^{\mathrm{MB}} Q^{\mathrm{MB}}(s_{t,1},m)}$$

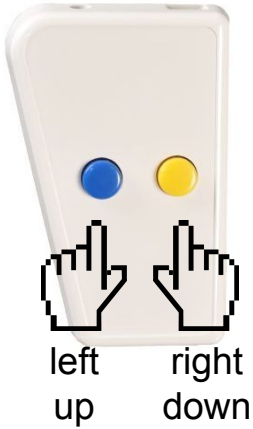
2-move trials:

$$\mathrm{p}(m_{t,1} = m|s_{t,1}) \propto e^{\gamma_m + \beta_2^{\mathrm{MF1}} Q_t^{\mathrm{MF1}}(s_{t,1},m) + \beta_2^{\mathrm{MF2}} Q_t^{\mathrm{MF2}}(s_{t,1},m) + \beta^{\mathrm{MB}} Q^{\mathrm{MB}}(s_{t,1},m)}$$

$$\mathrm{p}\left(m_{t,2} = m|s_{t,2}\right) \propto e^{\gamma_m + \beta_2^{\mathrm{MF1}} Q_t^{\mathrm{MF1}}\left(s_{t,2},m\right) + \beta_2^{\mathrm{MF2}} Q_t^{\mathrm{MF2}}\left(s_{t,1},m_{t,1},m\right) + \beta^{\mathrm{MB}} Q^{\mathrm{MB}}\left(s_{t,2},m\right)}$$

METHOD **MEG**

Why MEG?

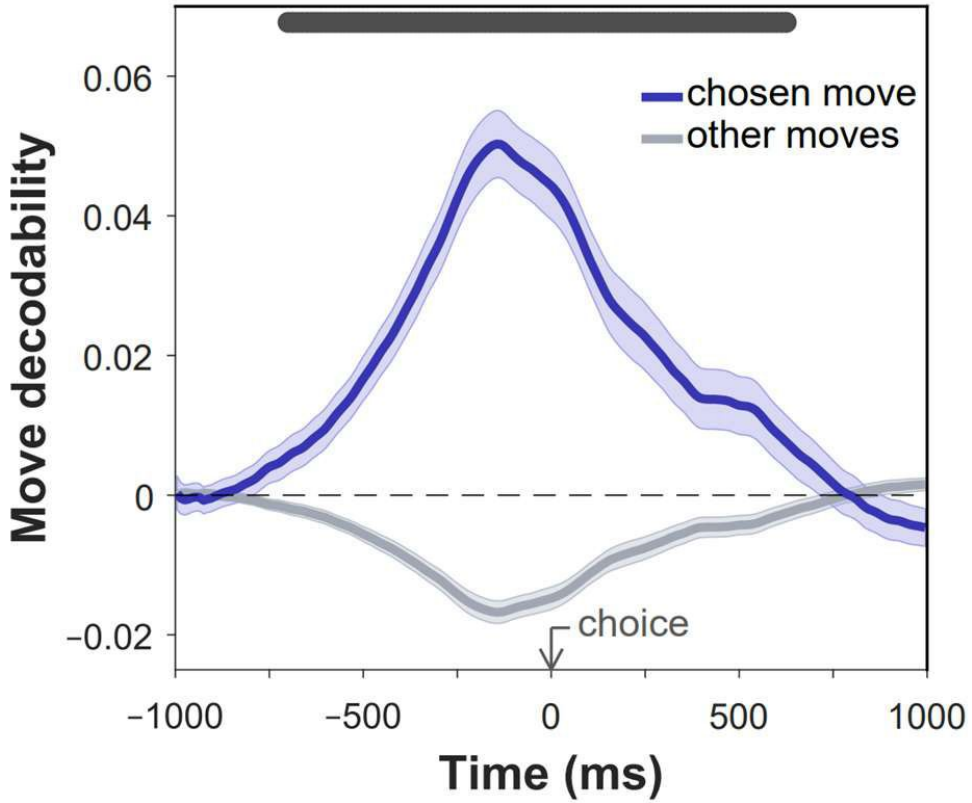


Training R F F F R N F F **Reward** R N F F R N F F **Spatial** R N F F

1 2 3 4 5

METHOD **MEG**

Move decoding



METHOD **MEG**

Image decoding

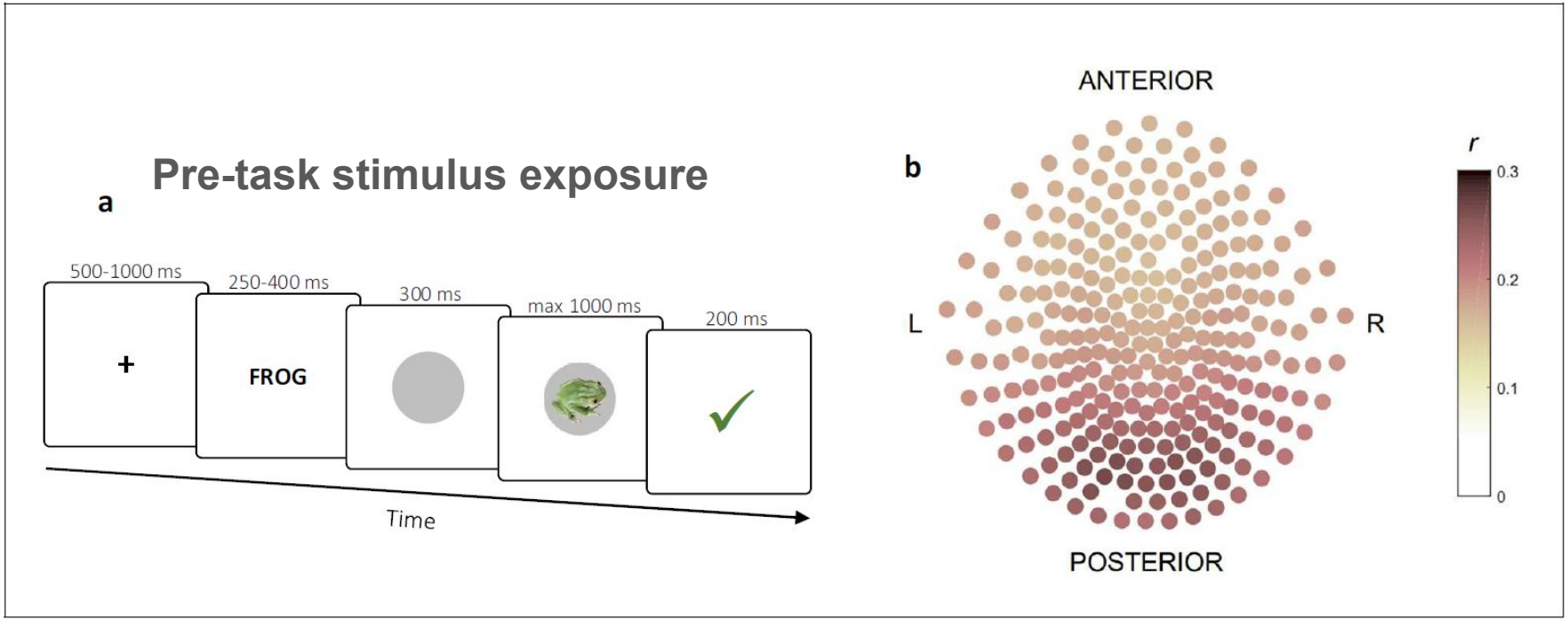
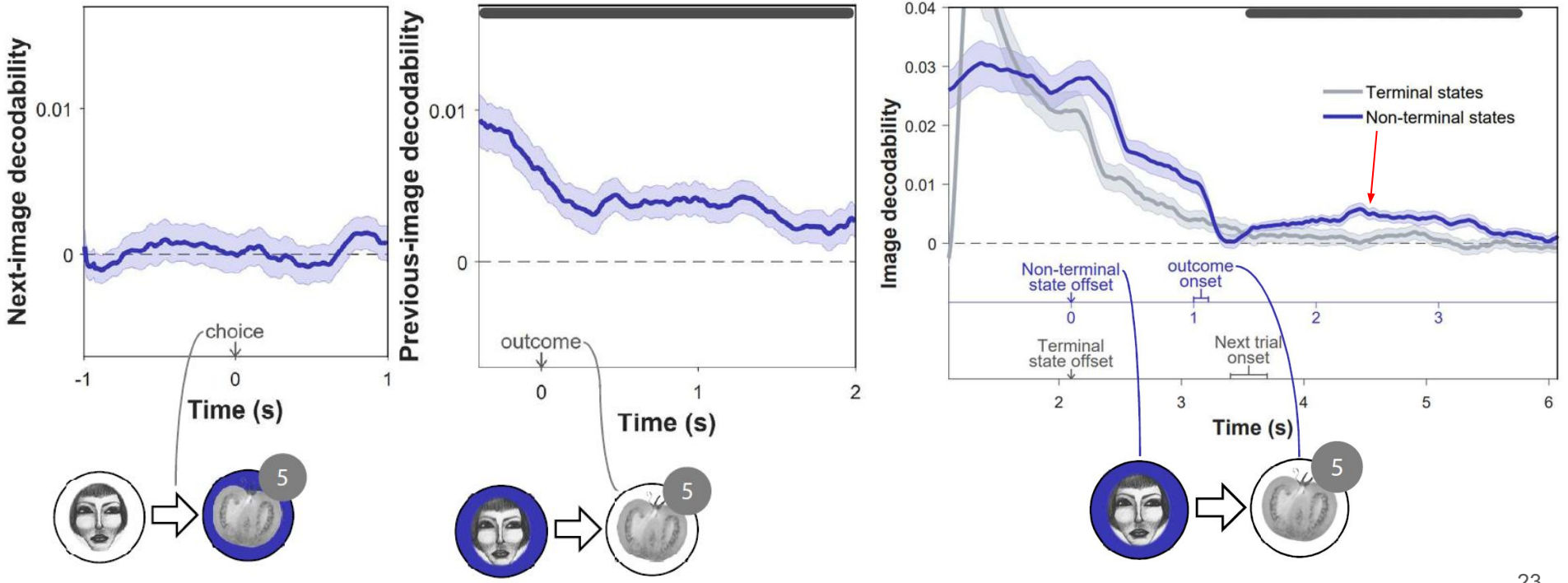


Figure 2—figure supplement 1. Decoding procedure. (a) Pre-task stimulus exposure on which decoders were trained. Timeline of a trial. (b) Decoding contribution by sensor. Contribution was quantified as the Spearman correlation between MEG signal and decoder output within trials for each stimulus. Correlations were then averaged over stimuli, trials and subjects.

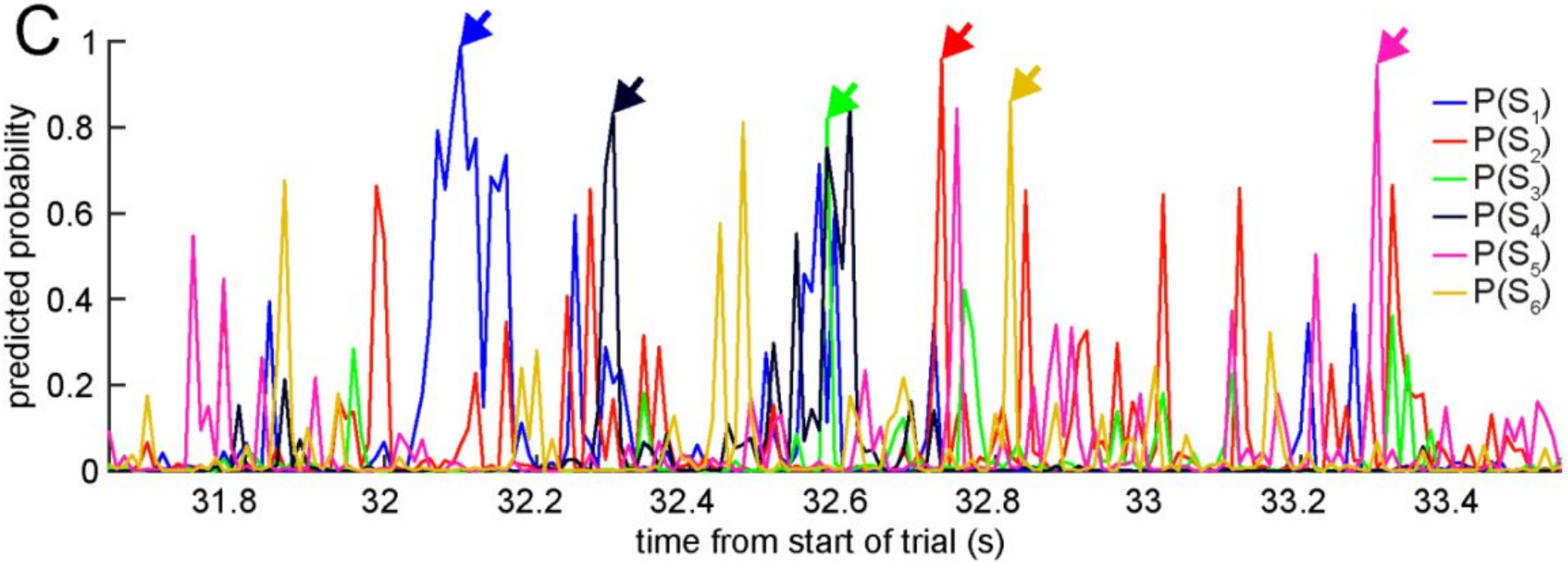
METHOD **MEG**

Previous, not subsequent, states were encoded in MEG



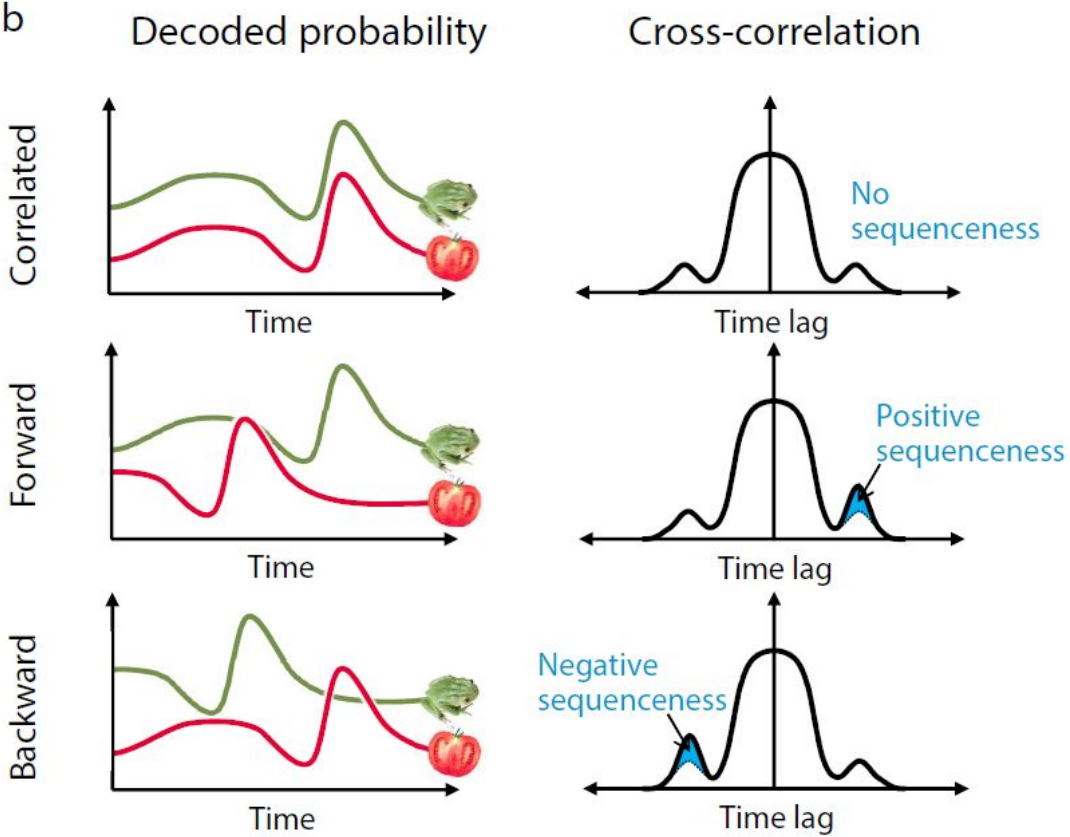
METHOD **Sequenceness**

Capture representation transitions



Kurth-Nelson et al. 2016

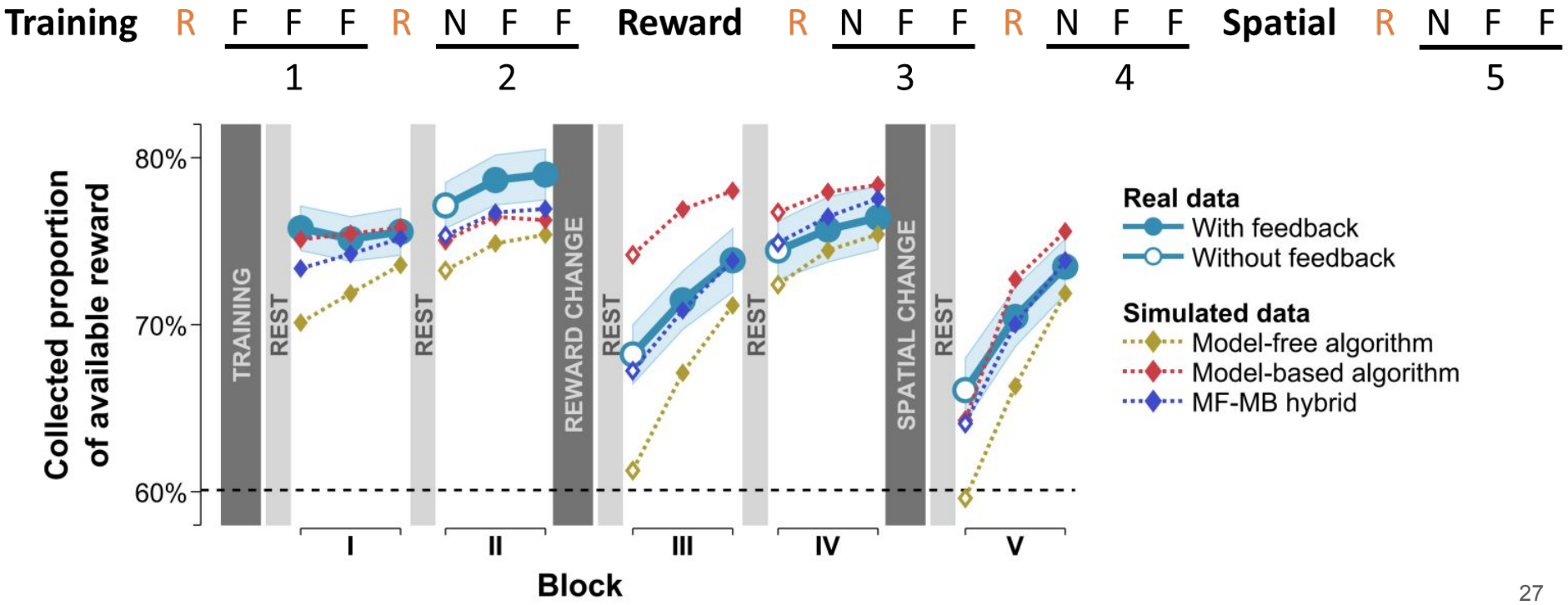
METHOD **Sequenceness**



5 MIN BREAK

RESULTS **Behavior**

Individual Differences

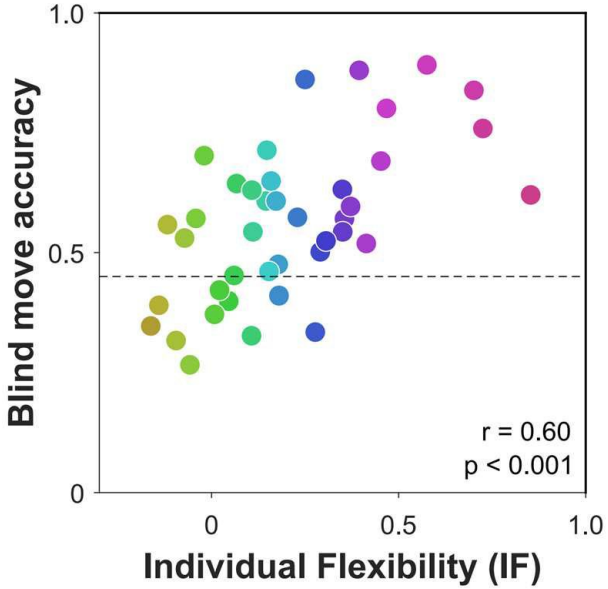


RESULTS **Behavior**

Planning two steps into the future

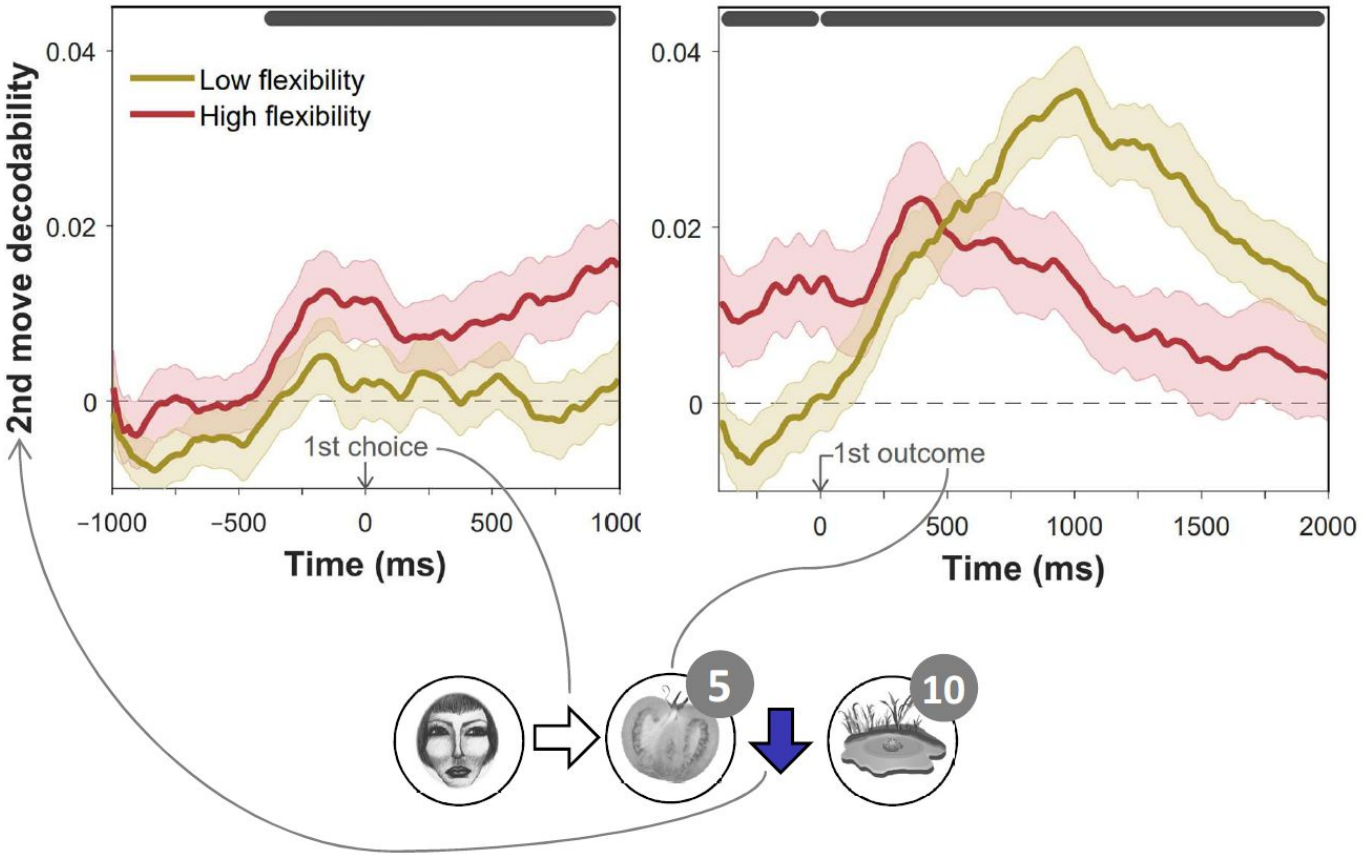
Training R F F F R N F F **Reward** R N F F R N F F **Spatial** R N F F

1 2 3 4 5



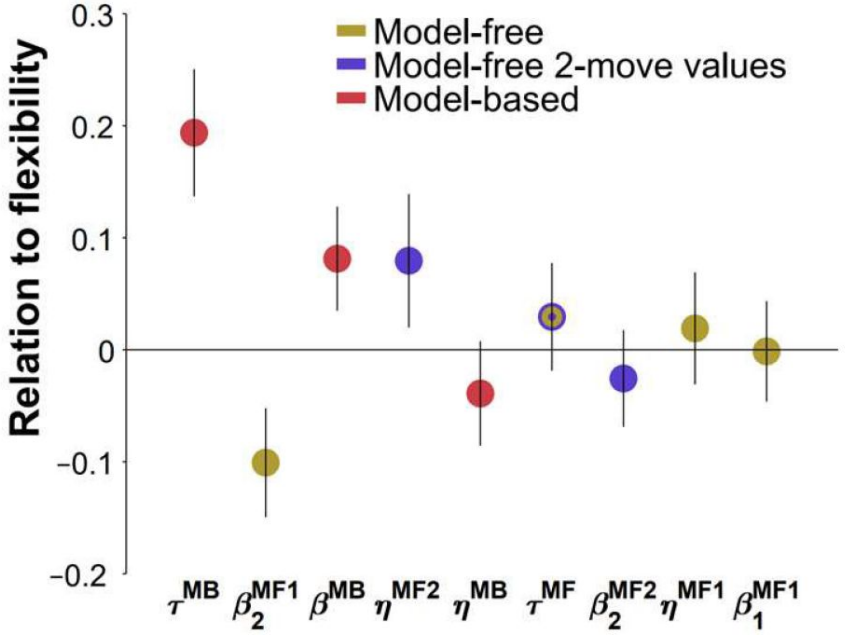
RESULTS **Behavior**

Planning two steps into the future



RESULTS **Behavior**

Individual flexibility reflected MF-MB balance

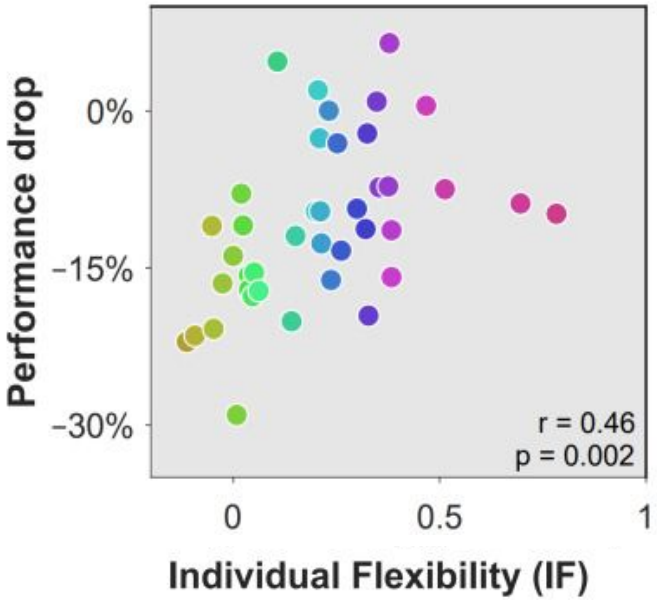


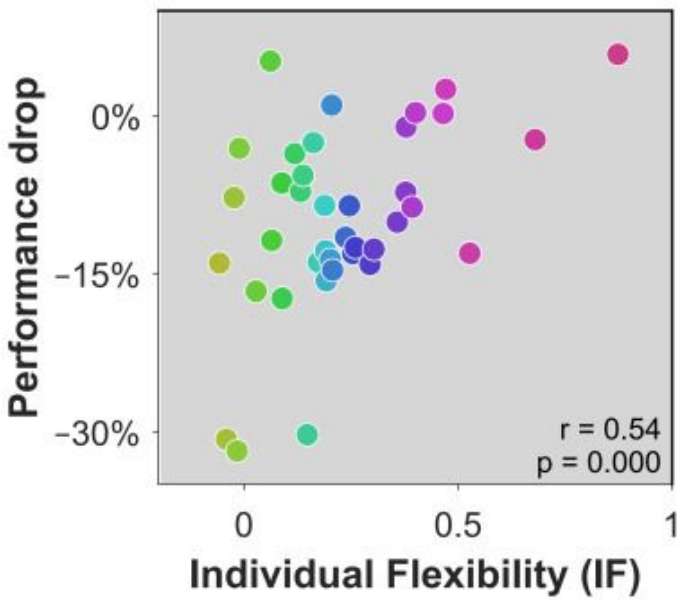
RESULTS **Behavior**

Individual flexibility reflected MF-MB balance

Training R F F F R N F F **Reward** R N F F R N F F **Spatial** R N F F

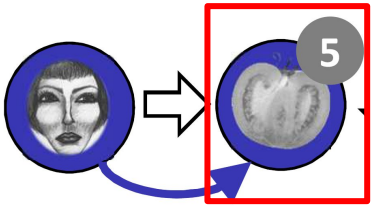
1 2 3 4 5

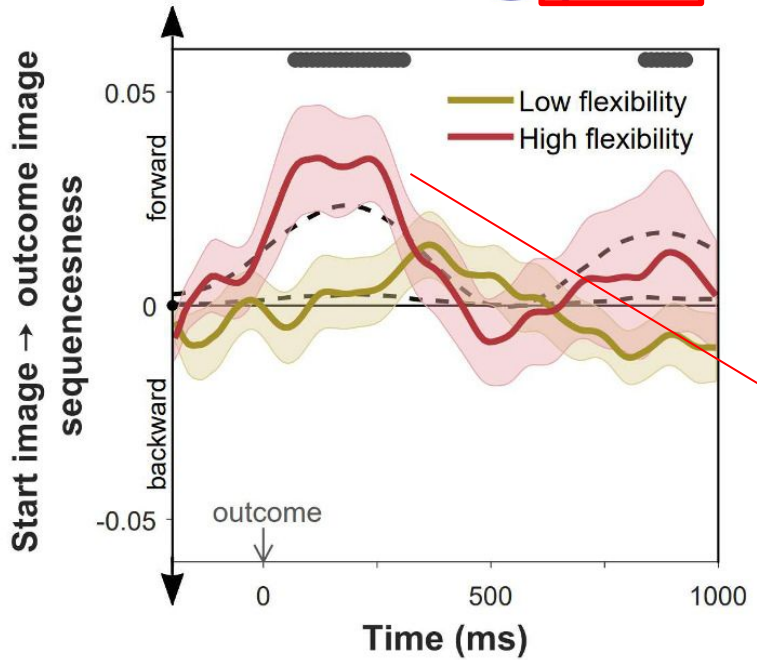


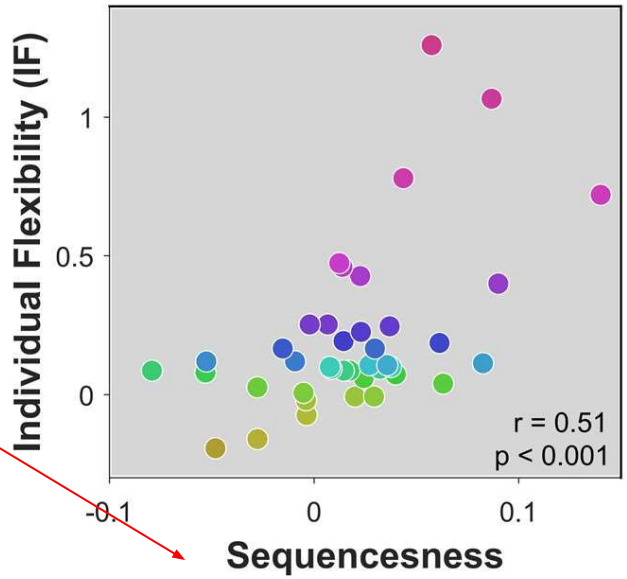


RESULTS ON Replay

is induced by prediction errors and associated with flexibility

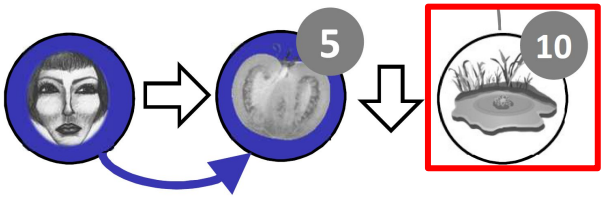


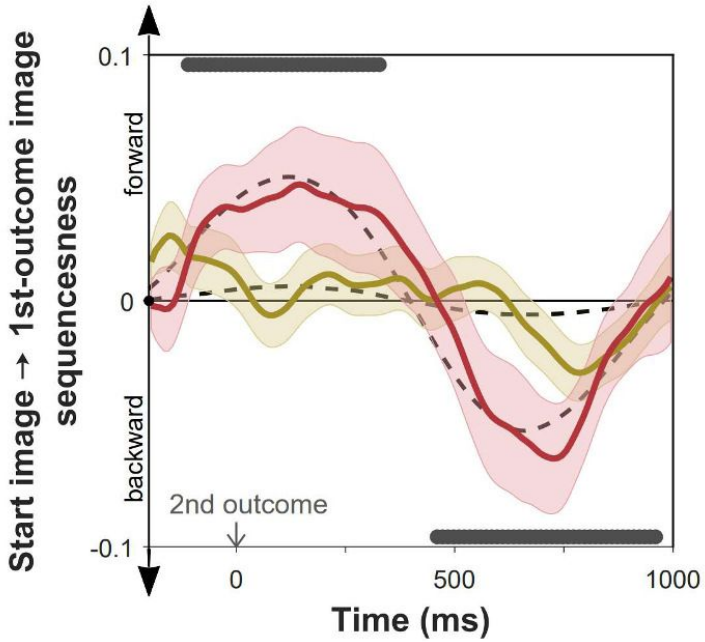


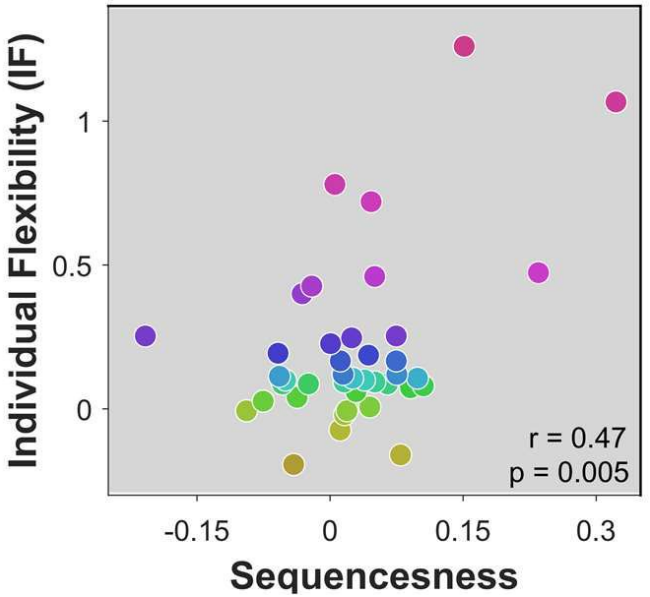


RESULTS ON Replay

is induced by prediction errors and associated with flexibility

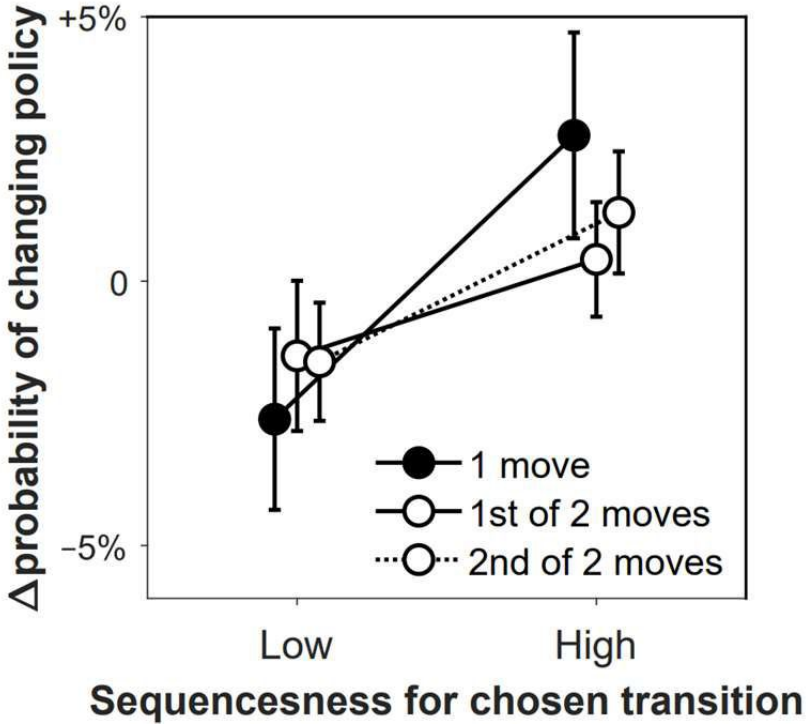






RESULTS **ON Replay**

is associated with policy update



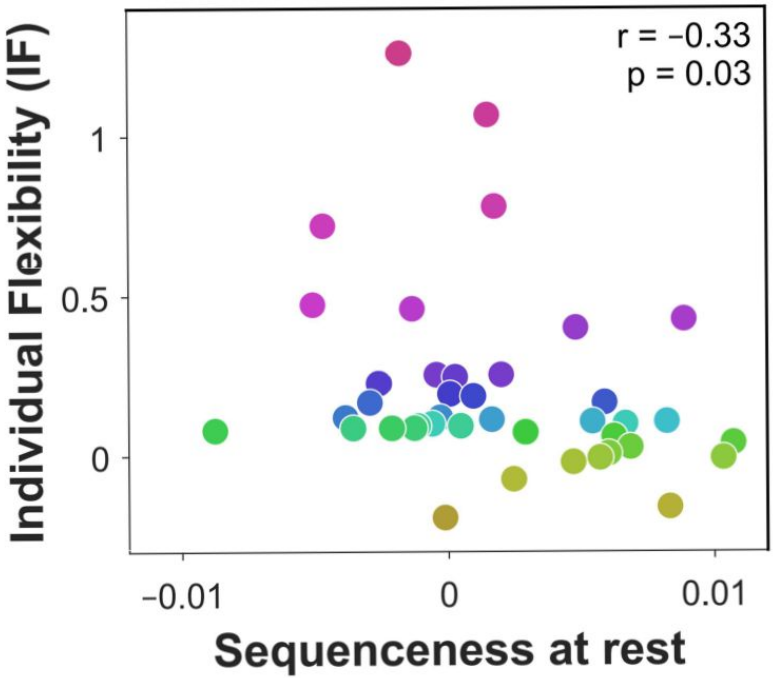
Choice re-evaluation
Replayed trajectories subsequently avoided

Post-outcome model-based planning?

RESULTS **OFF Replay** is induced by prediction errors and associated with low flexibility

Training R F F F R N F F **Reward** R N F F R N F F **Spatial** R N F F

1 2 3 4 5



Establish model-free policies?

RESULTS **OFF Replay**

Off-task replay can predict subsequently chosen sequences

Evidence of preplay - requires a model

OFF Replay negatively correlated with **IF**
OFF <u>Pre</u>play did not correlated with **IF**

Replay → less flexible
Planning → not change behavior systematically

"offline planning is ill-suited for enhancing trial-to-trial flexibility"

Summary

(1) inter-individual differences in the behaviour & flexibility in a simple state-based sequential decision-making task

(2) higher flexibility relies on better planning of future steps; with the second move being decodable already during the first-move choice

(3) flexibility is influenced more by MB compared to MF planning

(4) online replay is coupled to advantageous re-evaluation of choices

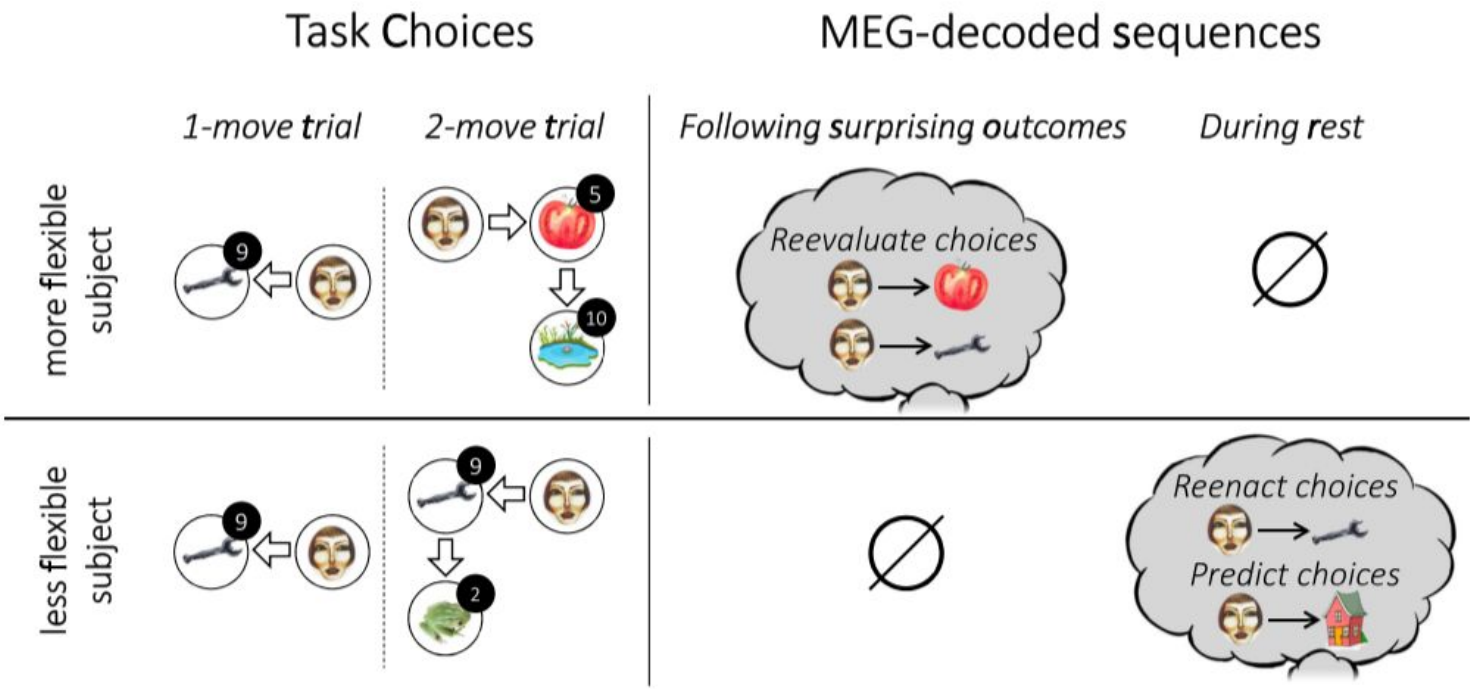
(5) offline replay (“prospective offline replay”) is involved in planning

Conclusions

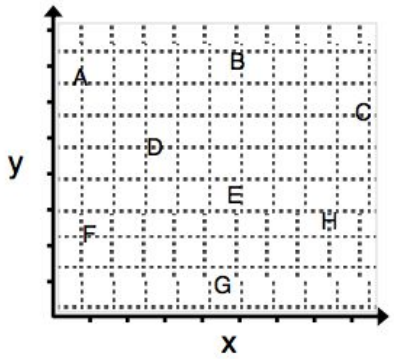
PREPLAY: forward replay should prioritise frequently chosen trajectories (those on which the agent might soon re-embark)

REPLAY: backward replay should prioritise trajectories for which one's policy can be informed, improved, and changed

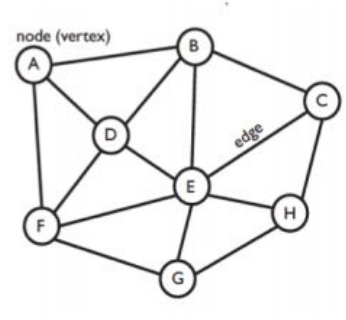
Graphical summary



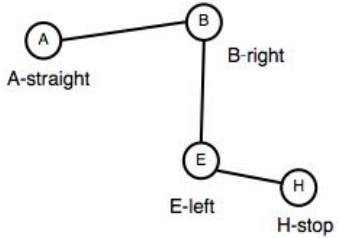
Which cognitive map?



- Euclidean map
  - » coordinate system
  - » metric structure
  - » distances & angles
- Supports:
  - » familiar routes
  - » novel detours
  - » novel shortcuts



- Topological Graph
  - » coordinate-free
  - » non-metric
  - » connectivity
- Supports:
  - » familiar routes
  - » novel detours
  - » NOT novel shortcuts



- Route
  - » place-action associations
  - » sequences
- Supports:
  - » familiar routes
  - » NOT intersecting routes
  - » NOT novel detours
  - » NOT novel shortcuts

Limitations of this study

(1) **Limited decision flexibility captured**
→ here only interleaving of 1-move and 2-move trials

(2) **Sequenceness measure**
→ focus on relative predominance of forward or backward; what about co-existence?

(3) **Image vs. state representation**
→ MEG decoding based on images versus concepts

(4) **Hippocampal vs. cortical replay**
→ MEG measures activity in cortical structures; what about hippocampal replay specifically?

Open questions

When is policy being updated? When information is available, prior to choice?

What is the cost of updating policy? What is the benefit?

What are the rules governing the choice of sequences to be replayed?

What about the correlation to behavioural disorders like OCD?

…

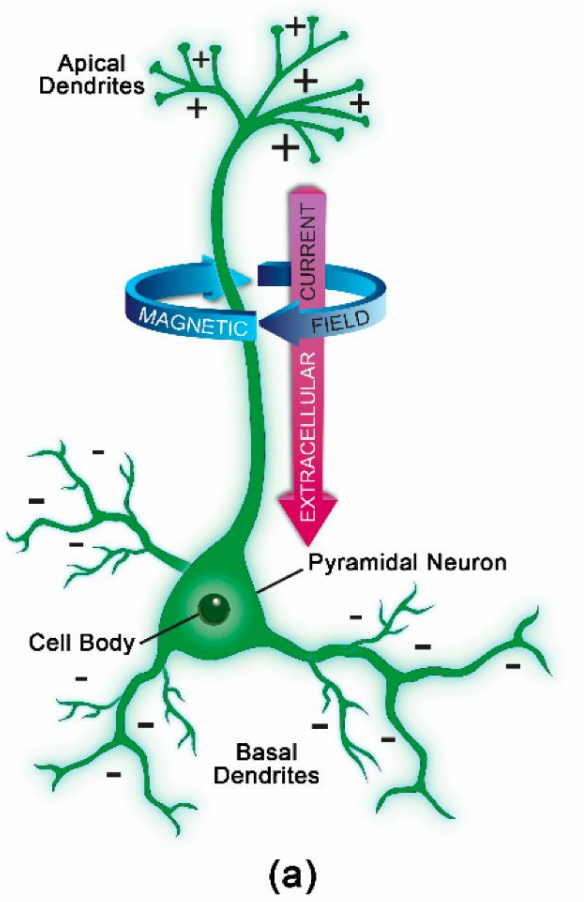
Discussion
Questions from you

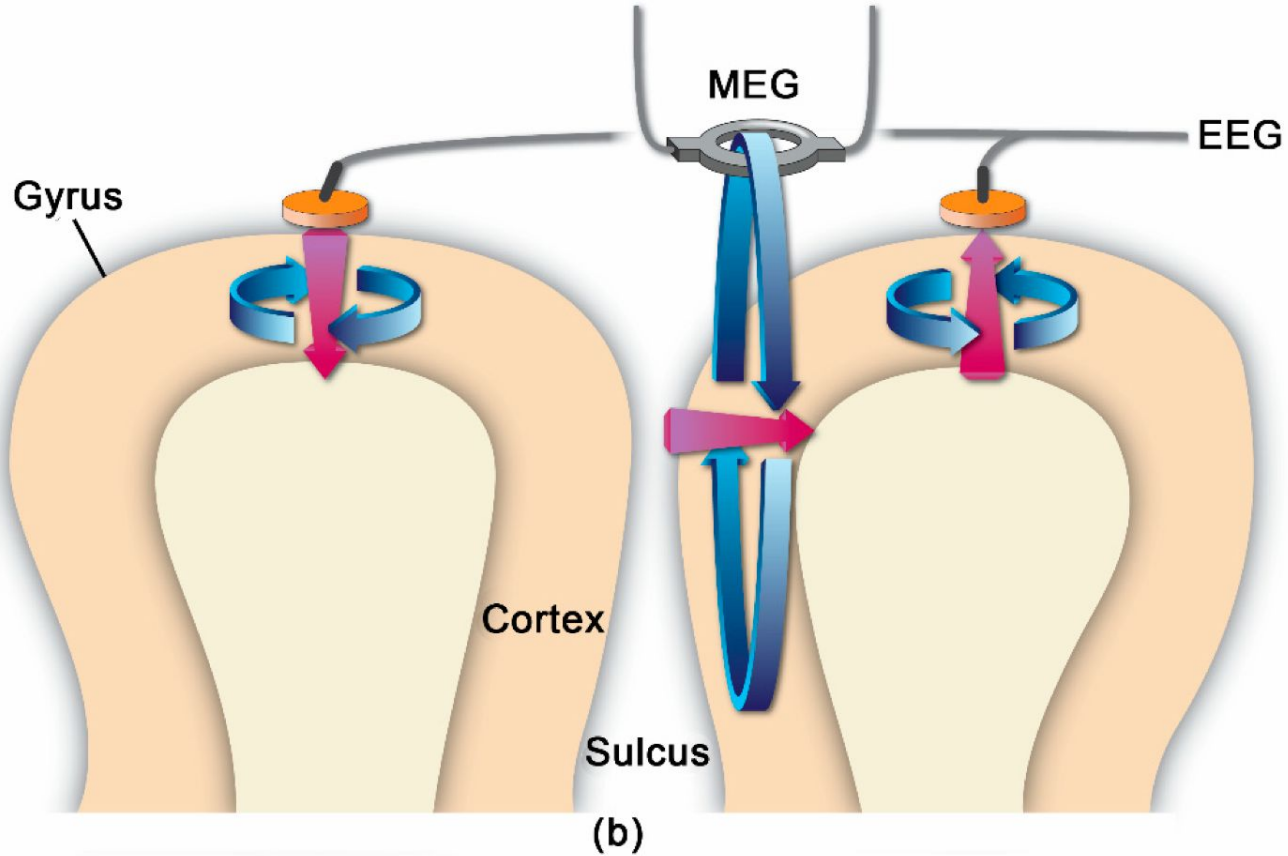
Q: Is replay only happening in structures related to planning and memory (e.g. hippocampus and prefrontal cortex) or everywhere in the brain...?
A: Replay originates in the hippocampus and can propagate to the cortex.

Q: How can we use MEG data to tell if someone is replaying a specific scenario and not just showing increased brain activity?
A: We can compare the activity of when people had to replay the specific scenario against control conditions (i.e., replaying another scenario). Increase activity is supposed to be observed in both conditions, but the pattern of activity would be different.

Q: Do “on-taskers” and “off-taskers” display different character traits by which one can recognize them?
A: Spontaneity? Rigidity? Rumination in depression? Link to disorders!

Q: Where specifically in the brain are these measurements done?





Discussion
Questions from you

Q: Does the experimental setting favor on-task replays strength?
A: “The task effectively incentivises flexible model-based reasoning, as this type of reasoning allows collection of substantial additional reward (93%) compared to our most successful MF algorithm (80%).”

Q: Do you think you can train yourself to become more flexible and thus become more of an online-replayer? Does the degree of flexibility (i.e. your usage of replay) remain stable with older age?
A: Maybe training is possible? How much is intrinsically anchored and how much is adjustable?

Q: Is there a switch between model-free and model-based planning?
A: Possibly.

Q: Is replay a purely unconscious thing, or is it possible that you are aware that you currently do it, say e.g. in "offline shower-thoughts"?
A: Replay can be induced (i.e. in memory retrieval), unclear whether this is “conscious”?

Discussion

Questions from you

Q: This experiment was only done during a rather short timespan between the tasks. Would the role of forward and backward replay differ if there was more time between the tasks?
A: I think it's reasonable to assume the role would be different. There are some studies with animals showing different conclusions, for example.

Q: Which policy can be expected to be dominant over longer timespans? That is, do we "forget" the policy that we developed during offline replay over time or do we repeat offline replay until we meet the same task again?
A: I believe it will depend on the task and what kind of strategy has the best cost-benefit, both in terms of mental effort but also in physiological terms.

Discussion

Questions from you

Q: In their model, they use prediction error/surprise as a predictor for sequenceness. For surprise, were they just looking at when the values of the images were changed? Or is there a feature like the P300 in EEG that they were using to detect surprise?
A: They used PE from the hybrid model.

Q: Also the idea of sequenceness/replay and reminded me of the Buzsaki paper about the HC as a sequencer… Is it possible the idea of flexibility/model-based corresponds to the ability to manipulate connected sequences? Reversing them to retrace causality, adding them together to build comprehensive and useful state space, and segmenting them according to the task at hand?
A: Possibly?

Discussion

Questions from you

Q: What could be the neural implementation (what exact neural circuits) that explains the different effects of Off-Line and On-Line replay?
A: Just a really general idea might be that MB is more prefrontal dependent, while MF more temporal dependent.

Q: How can replay be modulated such that it leads to more flexibility in one context (during the task: online), while it leads to less flexibility during another context (pause: offline).
A: Online and offline replay can lead to different effects even if the same neural circuits are involved. The mechanisms of information processing can still be different.

Q: Are they neuromodulatory systems (like dopamine) involved in this process?
A: We know that dopamine is involved in reward learning but I am not sure how much it is involved in the creation of a cognitive map… Anyone?

Discussion

Questions from you

Q: Could it be that flexibility and use of replay type of an individuum (flexible or not flexible) are connected by a third parameter like performance pressure, ambition driven character, fear of failure or insecurity when being observed (or even anxiety, OCD and other psychological phenomena)?
A: Sure, there are some studies showing a correlation of affective disorders and the tendency to use MF/MB in decision-making tasks.

Q: What could possibly be the neural structural correlate (if any) of the "switch" between the two modes of replay? Or could this possibly be explained sufficently by properties of network dynamics? Adding to Franzi's question, how would hyperparameters such as the one she mentioned regulate the switch and how?
A: No idea! If anyone knows, please tell us! I think both are possible but I would bet less money on a specific structure.

Thank you for your attention!

Good luck for your exams

Sources

Bera, Krishn & Mandilwar, Yash & Surampudi, Bapi. (2019). Value-of-Information based Arbitration between Model-based and Model-free Control.

Drieu, Céline & Zugaro, Michaël. (2019). Hippocampal Sequences During Exploration: Mechanisms and Functions. Frontiers in Cellular Neuroscience. 13. 10.3389/fncel.2019.00232.

Piastra, M. C., Nüßing, A., Vorwerk, J., Clerc, M., Engwer, C., & Wolters, C. H. (2021). A comprehensive study on electroencephalography and magnetoencephalography sensitivity to cortical and subcortical sources. *Human brain mapping*, *42*(4), 978–992.

Kurth-Nelson, Z., Economides, M., Dolan, R. J., & Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. *Neuron*, *91*(1), 194-204.