Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice

INTERNATIONAL
BRAIN
LABORATORY

Rylan Schaeffer

NeurIPS 2020

Research Aims

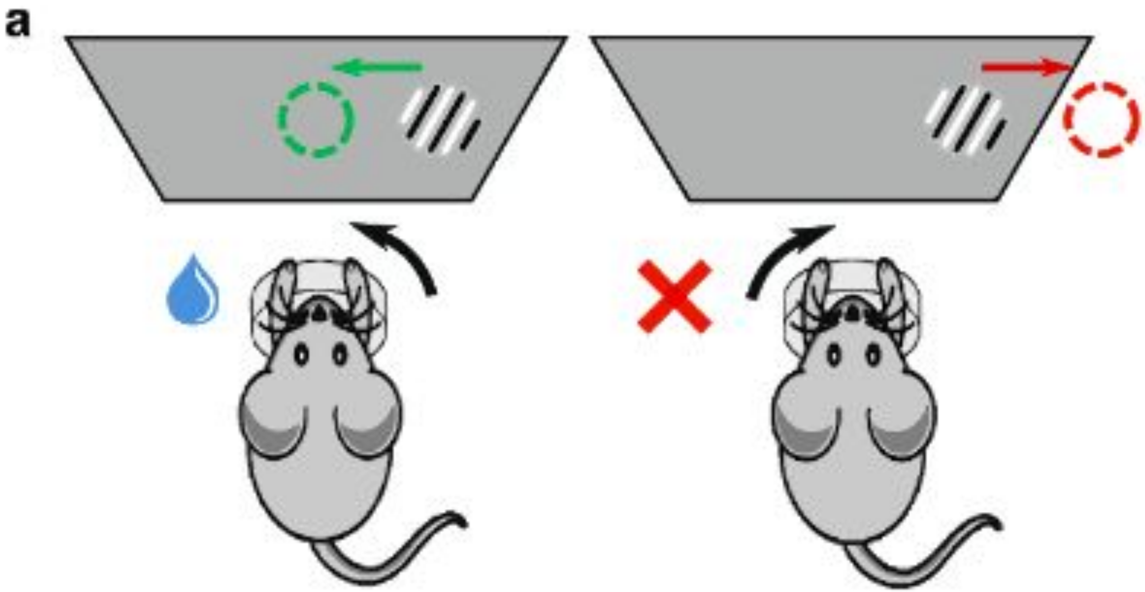
Goal: reverse engineer how recurrent neural networks perform hierarchical inference

Questions

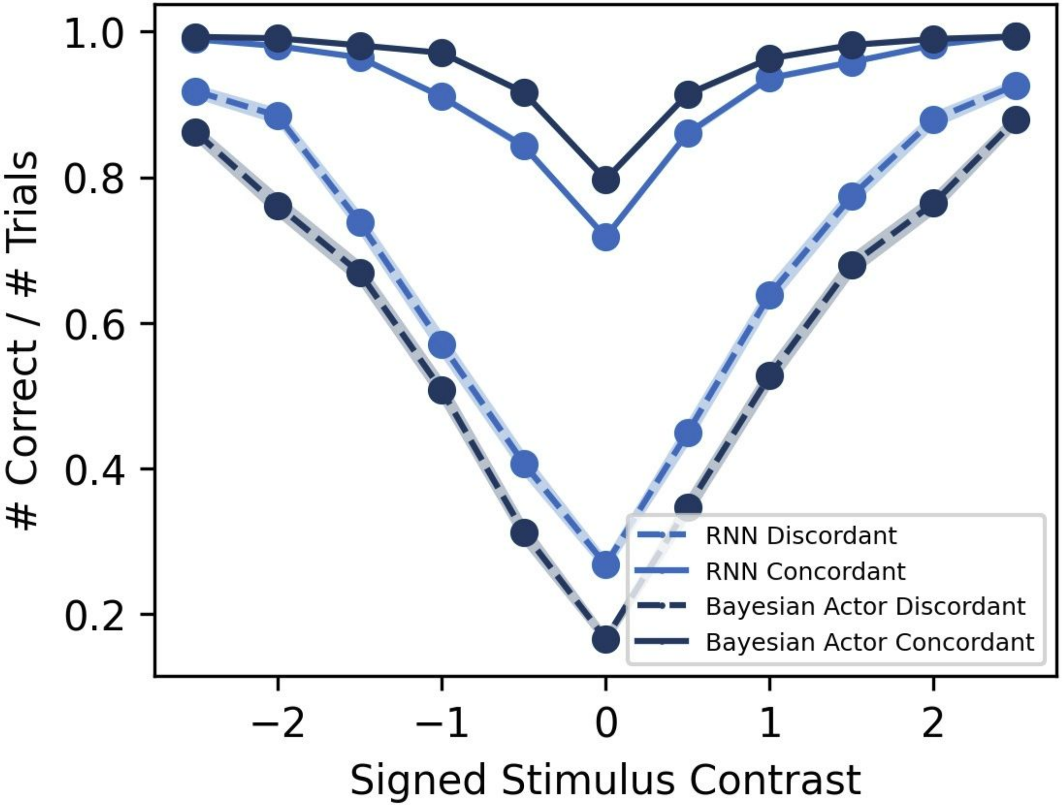
1. How well do RNNs compare against normative Bayesian baselines?

2. What are the representations, dynamics and mechanisms RNNs employ to perform inference?

Hierarchical Inference Task

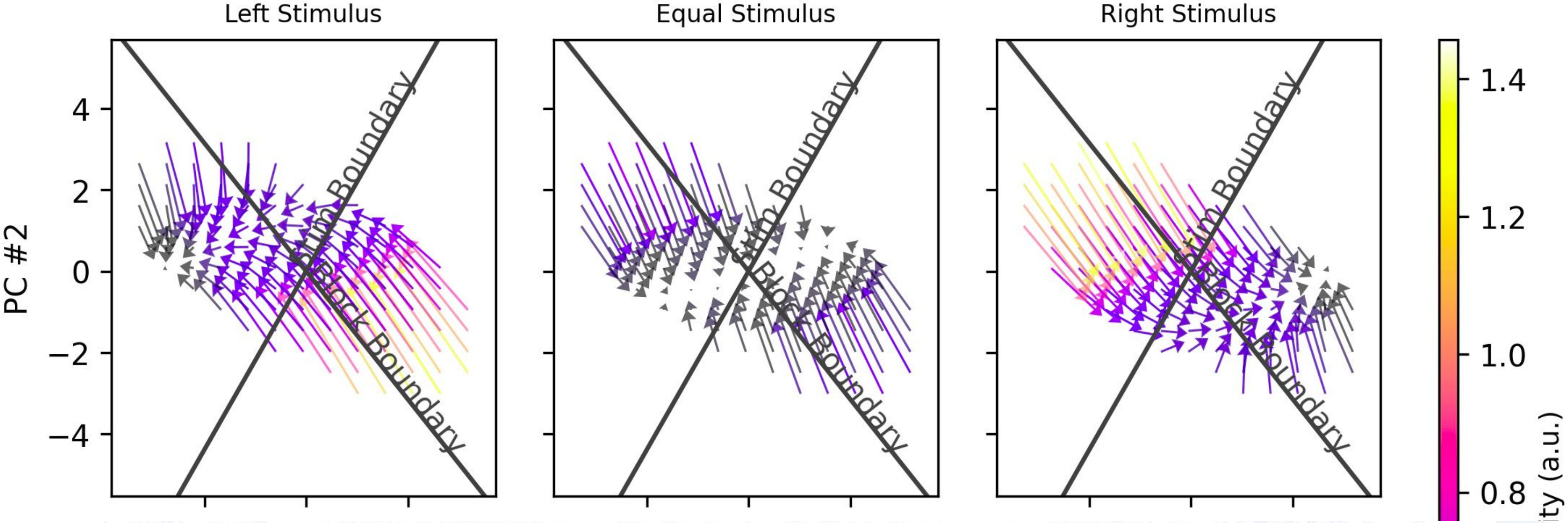


RNN Behavior Quantitatively Matches Bayesian Baseline



RNN State Space Displays Two Kinds of Dynamical Behavior

INTERNATIONAL
BRAIN
LABORATORY

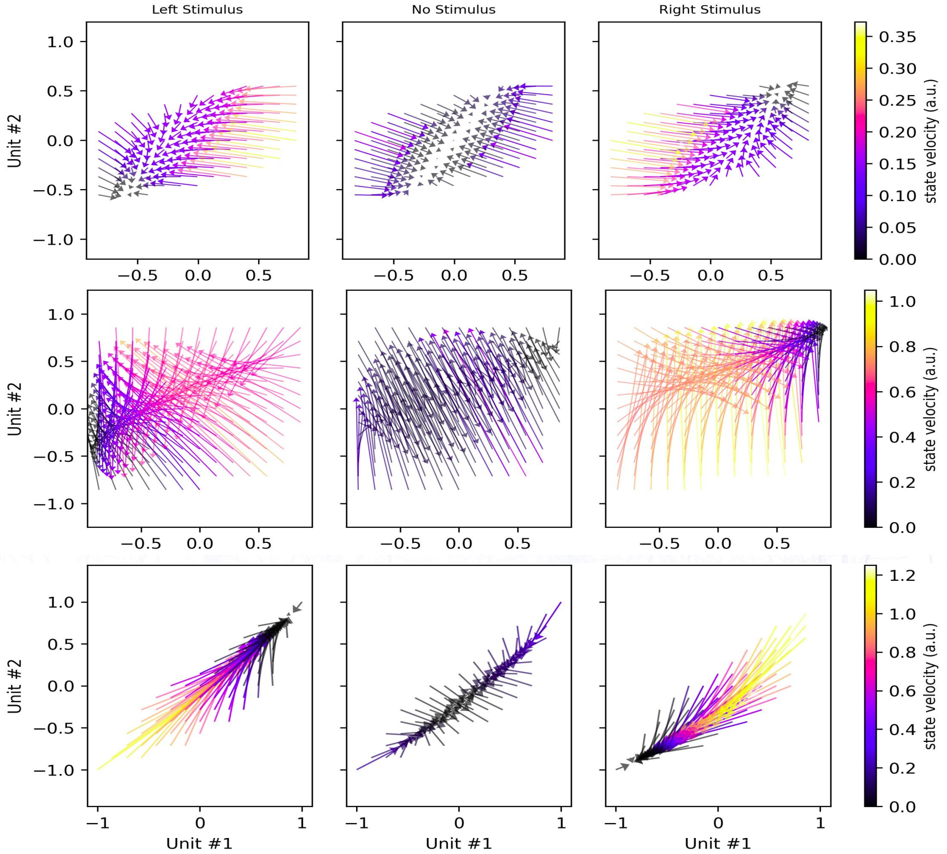


Novel Distillation Technique Preserves Phase Portrait

2-unit RADD RNN

2-unit Knowledge Distilled RNN

2-unit Task-Trained RNN



Novel Distillation Reveals RNN Circuit

$$\hat{z}_{n,t} = \begin{bmatrix} \text{Stimulus Belief}_{n,t} \\ \text{Block Belief}_{n,t} \end{bmatrix}$$

$$\hat{z}_{n,t} = \tanh\left(\begin{bmatrix} 0.54 & 0.31 \\ 0.19 & 0.84 \end{bmatrix} \hat{z}_{n,t-1} + \begin{bmatrix} -0.20 & 0.20 & 0.005 \\ -0.04 & 0.04 & 0.021 \end{bmatrix} \begin{bmatrix} o^L_{n,t} \\ o^R_{n,t} \\ r_{n,t} \end{bmatrix} \right)$$

Acknowledgements

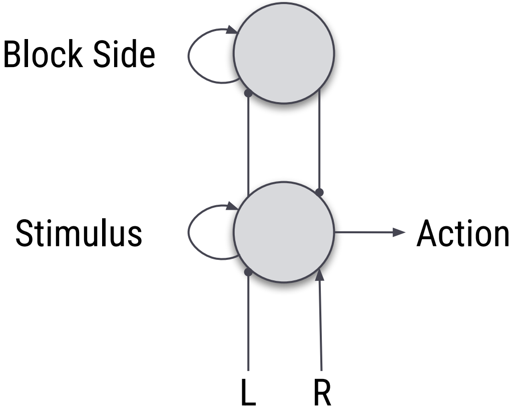
Rylan Schaeffer
Mikail Khona
Leenoy Meshulam
Ila Fiete
& IBL Theory Working Group

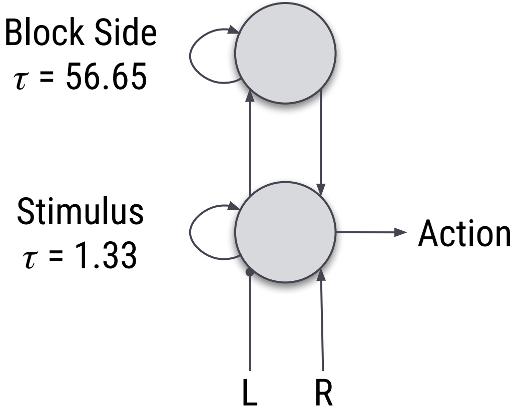
IACS Harvard John A. Paulson School of Engineering and Applied Sciences
IACS Institute for Applied Computational Science

McGOVERN INSTITUTE
FOR BRAIN RESEARCH AT MIT

brain+cognitive sciences

INTERNATIONAL
BRAIN
LABORATORY


Block Side
Stimulus
Action
L
R


Block Side
τ = 56.65
Stimulus
τ = 1.33
Action
L
R