

# The role of mental simulation in primate physical inference abilities

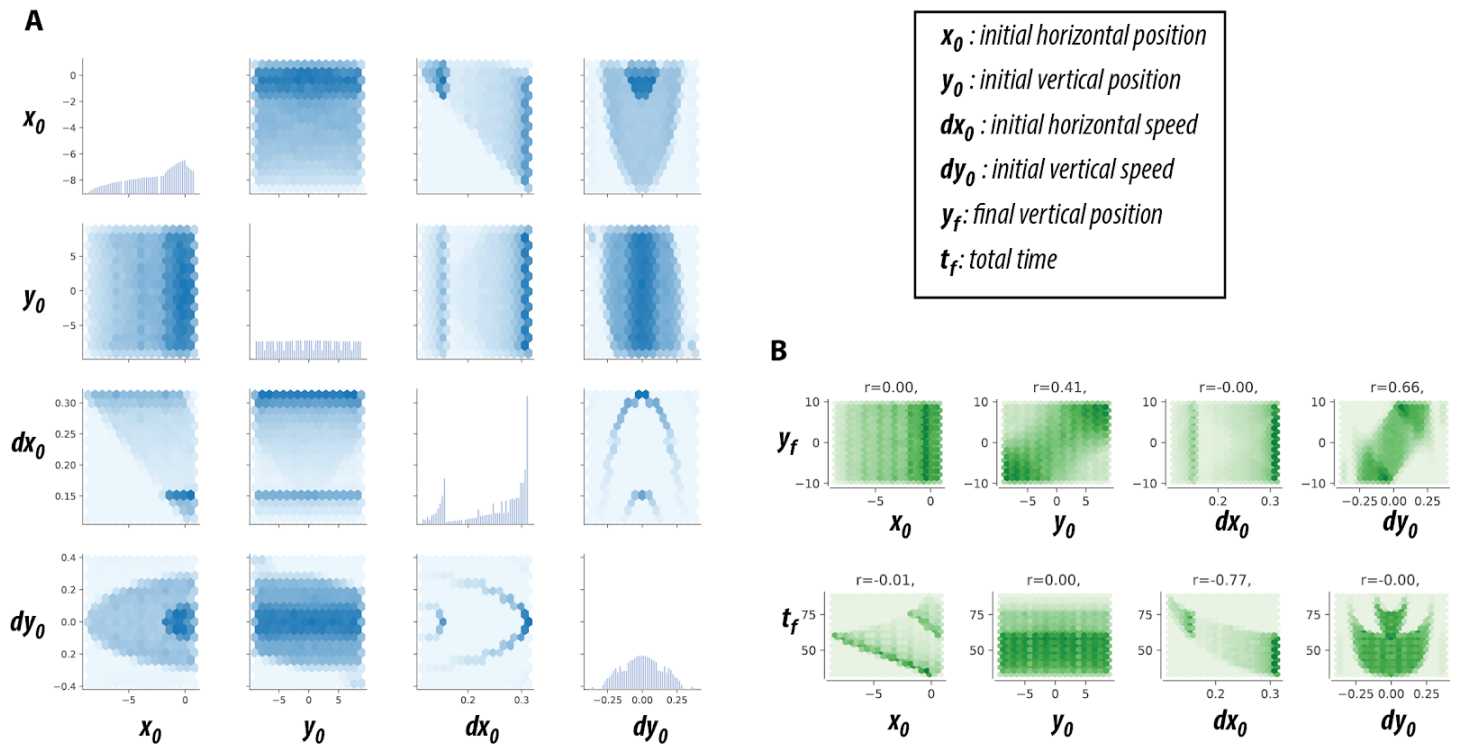
<sup>1</sup>Rishi Rajalingham, <sup>1</sup>Aida Piccato, and <sup>1,2</sup>Mehrdad Jazayeri

<sup>1</sup>McGovern Institute for Brain Research, <sup>2</sup>Department of Brain & Cognitive Sciences,

5 Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Correspondence: [mjaz@mit.edu](mailto:mjaz@mit.edu)

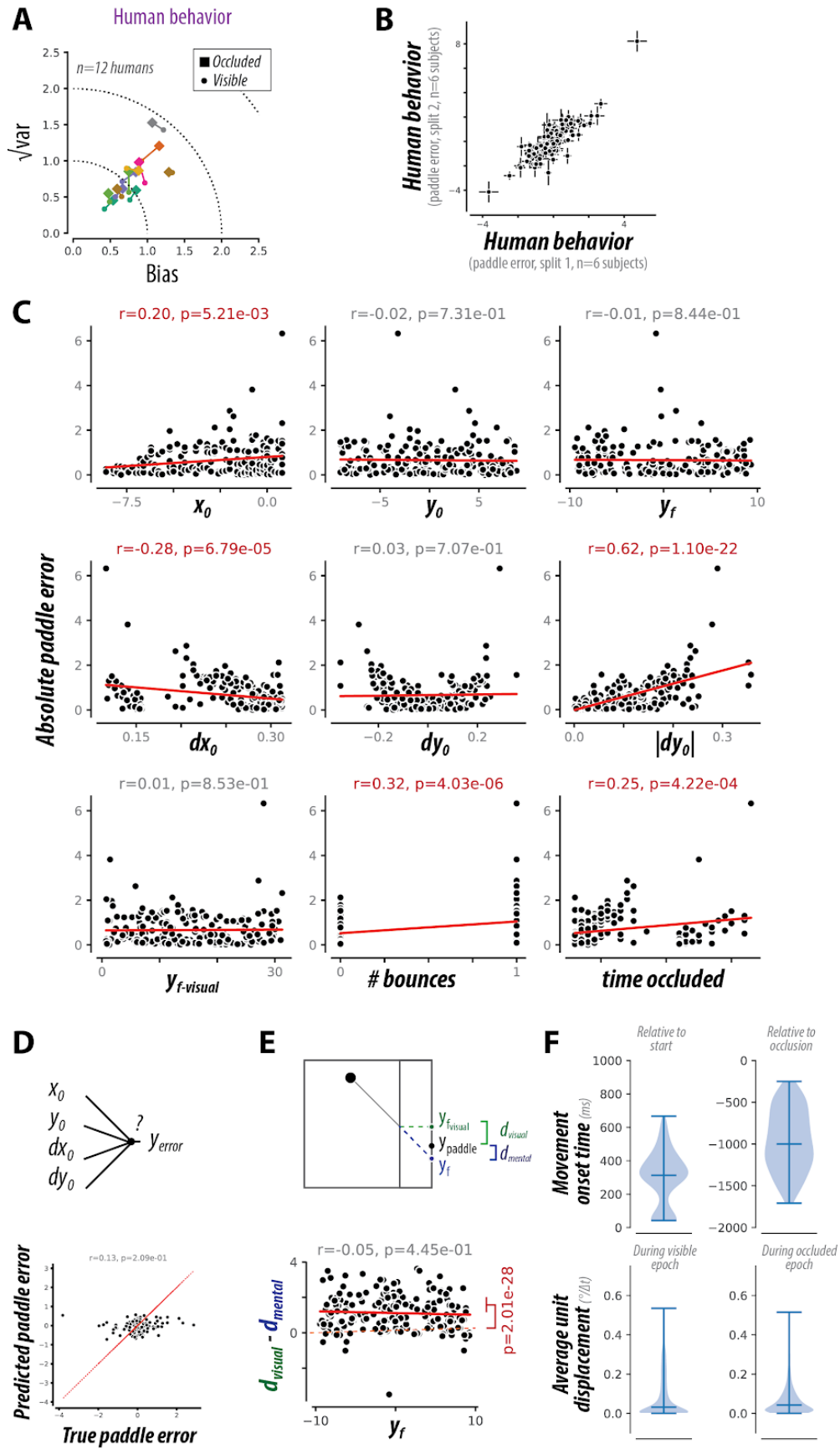
## Supplemental Information



**Figure S1. Task parameters. (A)** Distribution of meta-parameters corresponding to initial position ( $x_0, y_0$ ) and initial velocity ( $dx_0, dy_0$ ) for the task RNN training set. The diagonal panels correspond to histograms of each meta-parameter.

15 The off-diagonal panels correspond to 2D histograms showing the joint distribution for each pair of meta-parameters. **(B)** We sampled the initial ball position and velocity sufficiently broadly enough to ensure that the task was reasonably challenging. Each panel shows the joint distribution of each meta-parameter with the target output (final ball position,  $y_f$ , see top panels) and the time to interception ( $t_f$ , see bottom panels).

Figure S2:



25 **(A)** Human behavior for each individual subject, separately for visible and occluded trials. Recall that visible trials, which had a partially opaque occluder, were randomly interleaved with occluded trials. Performance is shown as a scatter of the bias (absolute error, after averaging across trials of the same condition) and variance (variability across trials of the same condition). The dotted circles correspond to lines of equal root-mean-squared error (RMSE). As expected, error is lower on visible trials.

30 **(B)** Comparison of behavioral error patterns between two spit-halves of human subjects reveals virtually identical error patterns, demonstrating that we measured a highly reliable human behavioral pattern. Trials from all 12 subjects were pooled to characterize the “archetypal” human behavior.

35 **(C)** Dependence of error patterns on task variables. Each panel shows absolute average endpoint error against task parameters varying per condition. The strength of each dependence, measured via a Pearson correlation, is shown on the corresponding panel titles; significant dependences are highlighted in red. Importantly, error patterns depended largely on dynamic variables relating to the ball speed ( $dx_0$ ,  $dy_0$ ,  $|dy_0|$ ,  $time\_occluded$ ) but not initial, intermediate, or final ball position ( $y_0$ ,  $y_{f\_visible}$ ,  $y_f$ ). Given the correlation between initial x position and ball speed built into this dataset (see Figure S1), errors were additionally correlated to the initial x position ( $x_0$ ). The occurrence of bounces also caused significantly greater errors, on average.

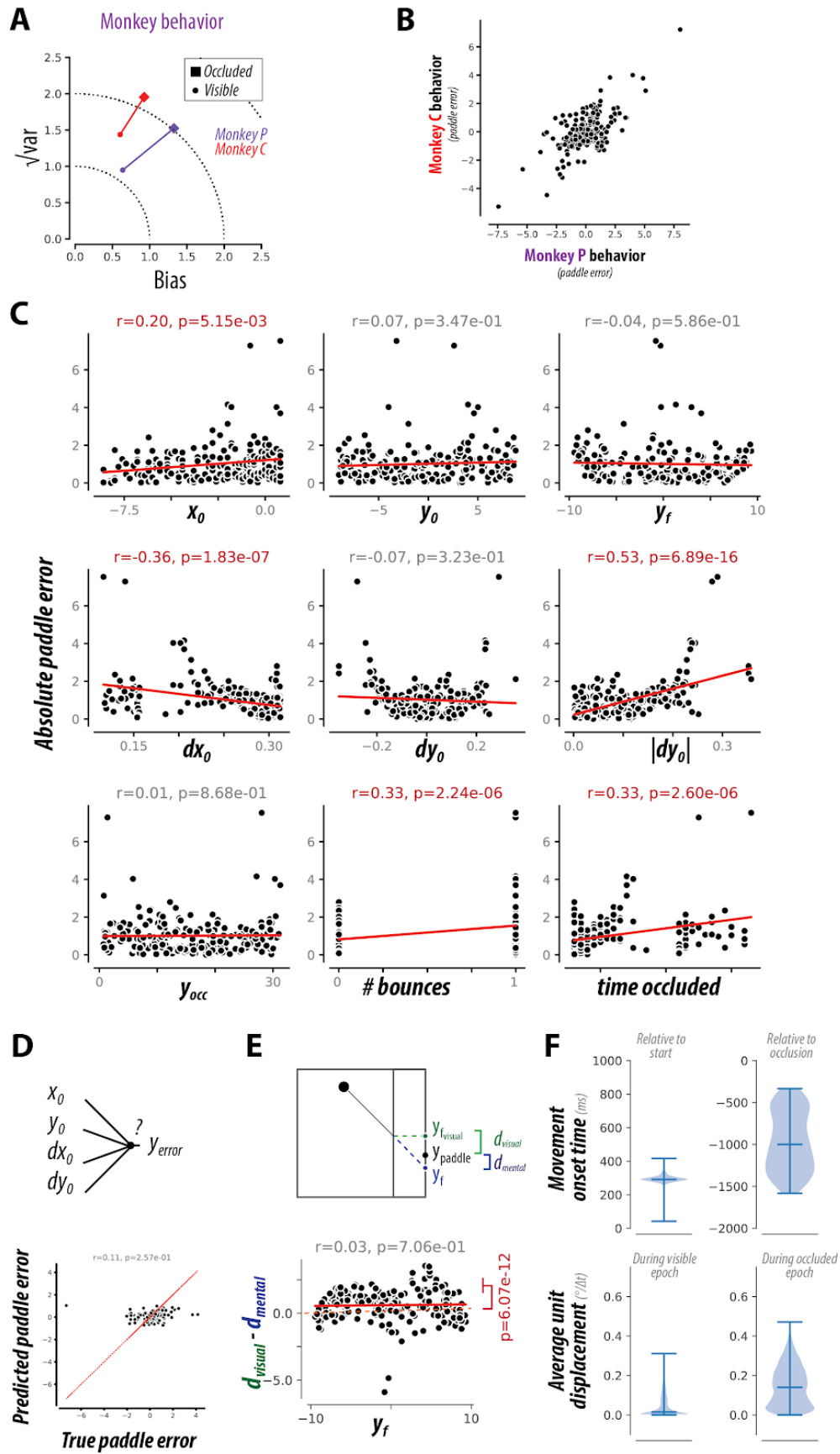
40 **(D)** Error patterns could not be explained by a simple function of initial position and velocity. We used cross-validated linear regression to predict the error pattern from the initial ball position and ball velocity; the resulting prediction was not better than expected by chance.

45 **(E)** Error patterns could not be explained by a simple visual tracking strategy, where the final paddle position is estimated based on the last visible position of the ball ( $y_{f\_visible}$ ). To demonstrate this, we compared the distance between the average final paddle position to each of the candidate final positions, as predicted by visual tracking ( $d_{visual}$ ) and mental tracking ( $d_{mental}$ ) strategies. We observe that the corresponding bias ( $d_{visual} - d_{mental}$ ) is significantly greater than zero (see red annotation) and moreover is not dependent on the final ball position (as might be expected if this bias is somehow driven by boundary conditions, see gray title annotation), demonstrating that behavioral error patterns are inconsistent with a simpler visual tracking strategy.

50 **(F)** Dynamics of movements. The top panels show the distribution of movement onset times, relative to the beginning of the trial (left) and to the beginning of the occluded epoch (right). Subjects initiated movements ~1000ms before the beginning of the occluded epoch. The bottom panels show the distribution of unit displacement during the visible and occluded epochs. Unit displacement was estimated from the instantaneous paddle position by first averaging across trials of the same conditions, and then measuring the mean absolute change in position. Subjects moved the paddle in both visible and occluded epochs, but displacements were significantly larger in the occluded epoch.

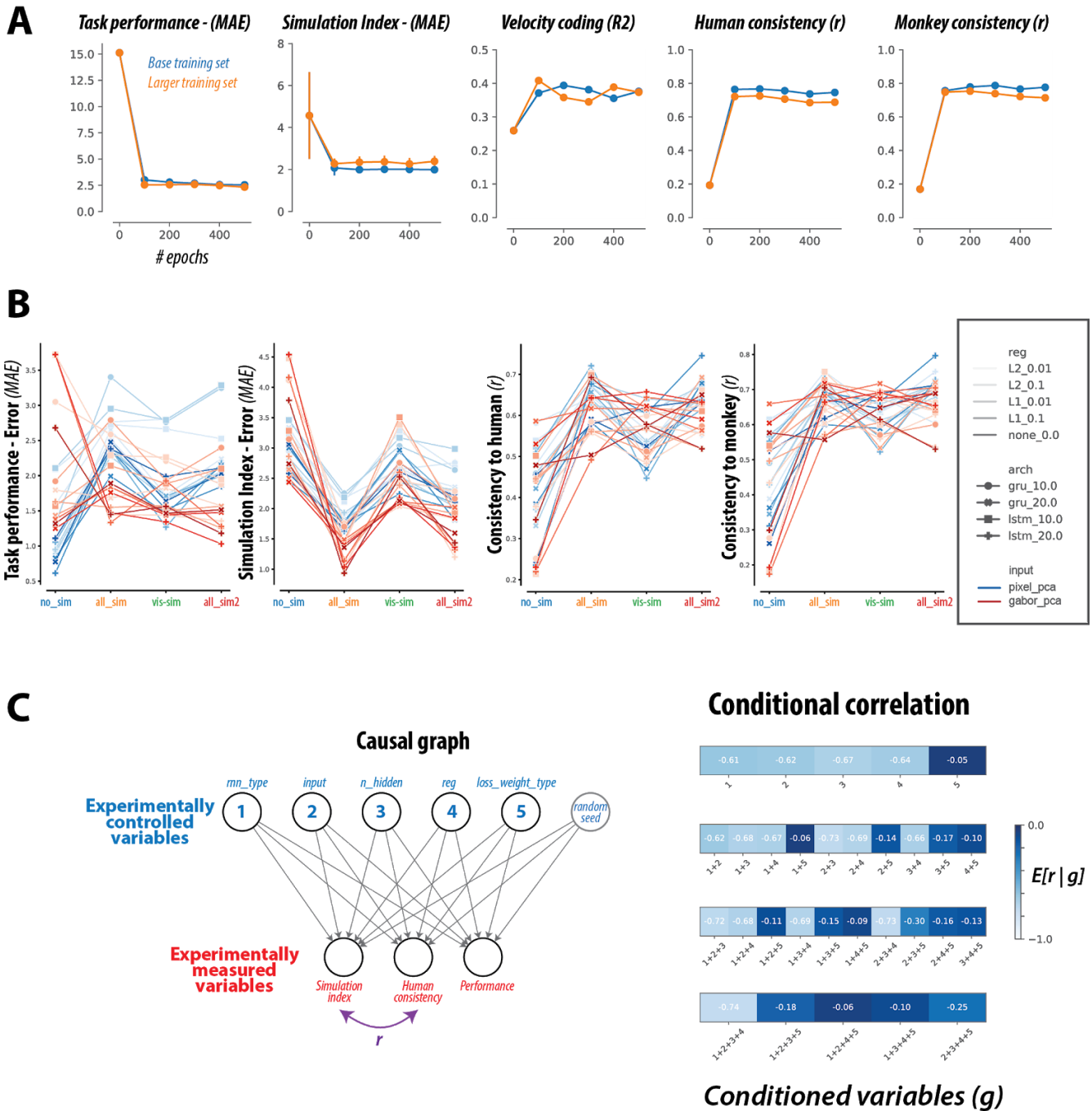
55

Figure S3:



(A-F) Formatting identical to Figure S2, but for monkey behavior.

Figure S4:



60

65

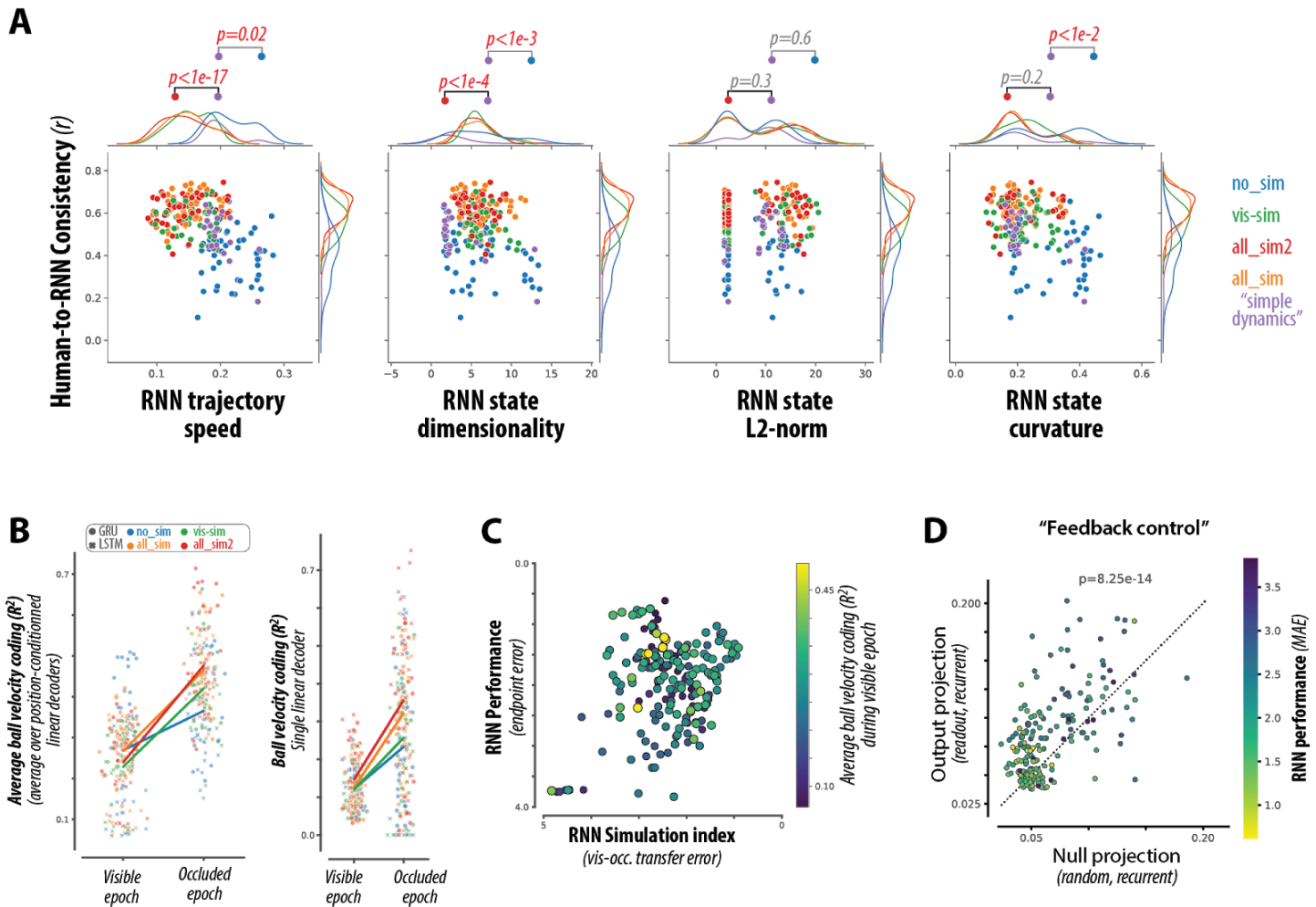
(A) Effect of RNN optimization. For the RNN model architecture with the highest human-consistency score, we evaluated key RNN metrics (e.g. performance, simulation index, consistency to humans and to monkeys) while varying both the number of training epochs and the training data (number of training samples and distribution of training data). We found that these metrics were largely insensitive to such variations in RNN optimization, suggesting that the extent of RNN training was sufficient to converge upon “stable” network solutions, and that our key results and inferences are largely robust of the details of this optimization procedure. (B) Different RNN models varied with respect to several hyper-parameters: different cell types (*rnn\_type*: LSTM or GRU), number of cells (*n\_hidden*: 10 or 20), input representation types (*input*: pixel\_pca or gabor\_pca), and regularization types (*reg*: L1\_0.01, L1\_0.1, L2\_0.01, or

70  $L2_{0.1}$ ); and were differently optimized (**loss\_weight\_type**: *no\_sim*, *vis\_sim*, *all\_sim*, or *all\_sim2*). Each of the four panels shows the effect of each hyperparameter choice on performance metrics (task performance, simulation index) and primate consistency (with respect to both human and monkey behavior). **(B)** The left panel shows the causal graph of our RNN experiments. We experimentally controlled five different RNN hyper-parameters (top row, blue), and from the resulting RNN model instances, we measured several attributes (bottom row, red), including the simulation index and the

75 human consistency. To uncover which RNN hyper-parameters *cause* the strong negative correlation ( $r$ ) between simulation index and human consistency, we measured the conditional correlation  $E[r|g]$ , conditioning on each of the 30 possible combinations ( $g$ ) of the five hyper-parameter types. The right panel shows the resulting conditional correlations for each of the 30 combinations. Darker values correspond to smaller magnitude correlations. The observed correlation between the simulation index and human consistency is largely driven by the hyper-parameter defining the optimization

80 target (**loss\_weight\_type**).

Figure S5:



85

(A) As a control, we optimized a new set of RNN models on task performance with additional regularization to promote simple dynamics, by adding regularization terms related to the L2-norms of the hidden state activity and the derivative of the hidden state activity (see Methods). To verify that this regularization had the intended effect on RNN representations, the four panels show the distribution of human-consistency against four representational metrics (dimensionality, speed, curvature, and norm) for all trained RNN models, grouped by their optimization type. Statistical comparisons between relevant distributions are shown above each scatter (unpaired t-test). The top comparison corresponds to 'no\_sim' vs 'simple\_dynamics' (blue vs purple), and the bottom comparison corresponds to 'all\_sim2' vs 'simple\_dynamics' (red vs purple). We observe that for all metrics that differed across RNN types, the 'simple\_dynamics' RNNs (purple) did indeed diverge from the baseline 'no\_sim' (blue) models as intended. (B) Comparison of average velocity coding during visible and occluded epochs, for all RNN models. The left panel shows velocity coding estimated via position-conditioned linear decoders, whereas the right panel shows the corresponding estimates with a single position-independent linear decoder. (C) Average velocity coding during visible epoch is not correlated to either performance or simulation index, in contrast to the corresponding metric during the occluded epoch (see Figure 5E). (D) We estimated a measure of "feedback control" to characterize the alignment between the read-out weights and the recurrent weights. While RNNs do not receive explicit instantaneous visual feedback, this metric aims to capture the extent to which the output of the network is fed back into its activity. The scatter plot shows the comparison of this metric against the null, for each RNN model; marker color corresponds to model performance. We observed that the median amount of feedback control was significantly greater than expected by chance.

100





## References

- 110 Andalman, Aaron S., Vanessa M. Burns, Matthew Lovett-Barron, Michael Broxton, Ben Poole, Samuel J. Yang, Logan Grosenick, et al. 2019. "Neuronal Dynamics Regulating Brain and Behavioral State Transitions." *Cell* 177 (4): 970–85.e20.
- Battaglia, Peter W., Jessica B. Hamrick, and Joshua B. Tenenbaum. 2013. "Simulation as an Engine of Physical Scene Understanding." *Proceedings of the National Academy of Sciences* 110 (45): 18327–32.
- 115 Chaisangmongkon, Warasinee, Sruthi K. Swaminathan, David J. Freedman, and Xiao-Jing Wang. 2017. "Computing by Robust Transience: How the Fronto-Parietal Network Performs Sequential, Category-Based Decisions." *Neuron* 93 (6): 1504–17.e4.
- Collins, Jasmine, Jascha Sohl-Dickstein, and David Sussillo. 2016. "Capacity and Trainability in Recurrent Neural Networks." *arXiv [stat.ML]*. arXiv. <http://arxiv.org/abs/1611.09913>.
- 120 Craik, K. J. W. 1952. *The Nature of Explanation*. CUP Archive.
- DiCarlo, J. J., and K. O. Johnson. 1999. "Velocity Invariance of Receptive Field Structure in Somatosensory Cortical Area 3b of the Alert Monkey." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 19 (1): 401–19.
- Egger, Seth W., Evan D. Remington, Chia-Jung Chang, and Mehrdad Jazayeri. 2019. "Internal Models of Sensorimotor Integration Regulate Cortical Dynamics." *Nature Neuroscience* 22 (11): 1871–82.
- 125 Fischer, Jason, John G. Mikhael, Joshua B. Tenenbaum, and Nancy Kanwisher. 2016. "Functional Neuroanatomy of Intuitive Physical Inference." *Proceedings of the National Academy of Sciences* 113 (34): E5072–81.
- Funahashi, Ken-Ichi, and Yuichi Nakamura. 1993. "Approximation of Dynamical Systems by Continuous Time Recurrent Neural Networks." *Neural Networks*. [https://doi.org/10.1016/s0893-6080\(05\)80125-x](https://doi.org/10.1016/s0893-6080(05)80125-x).
- 130 Gao, Peiran, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. 2017. "A Theory of Multineuronal Dimensionality, Dynamics and Measurement." *bioRxiv*. <https://doi.org/10.1101/214262>.
- Gerstenberg, Tobias, and Joshua B. Tenenbaum. 2017. "Intuitive Theories." *Oxford Handbook of Causal Reasoning*, 515–48.
- 135 Golub, Matthew, and David Sussillo. 2018. "FixedPointFinder: A Tensorflow Toolbox for Identifying and Characterizing Fixed Points in Recurrent Neural Networks." *Journal of Open Source Software* 3 (31): 1003.
- Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Networks." *arXiv Preprint arXiv:1406.2661* 4 (5): 6.
- 140 Hammer, Barbara. 2000. "On the Approximation Capability of Recurrent Neural Networks." *Neurocomputing* 31 (1): 107–23.
- Hamrick, Jessica B. 2019. "Analogues of Mental Simulation and Imagination in Deep Learning." *Current Opinion in Behavioral Sciences* 29 (October): 8–16.
- 145 Hassabis, Demis, Dhharshan Kumaran, and Eleanor A. Maguire. 2007. "Using Imagination to Understand the Neural Basis of Episodic Memory." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 27 (52): 14365–74.
- Hegarty, Mary. 2004. "Mechanical Reasoning by Mental Simulation." *Trends in Cognitive Sciences* 8 (6): 280–85.
- 150 Hessel, Matteo, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. 2017. "Rainbow: Combining Improvements in Deep Reinforcement Learning." *arXiv [cs.AI]*. arXiv. <http://arxiv.org/abs/1710.02298>.
- Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework." <https://openreview.net/pdf?id=Sy2fzU9gl>.
- 155 Hoch, Stephen J. 1985. "Counterfactual Reasoning and Accuracy in Predicting Personal Events." *Journal of*

*Experimental Psychology. Learning, Memory, and Cognition* 11 (4): 719–31.

- 160 Johnson, Kenneth O., Steven S. Hsiao, and Takashi Yoshioka. 2002. “Neural Coding and the Basic Law of Psychophysics.” *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 8 (2): 111–21.
- Kanitscheider, Ingmar, and Ila Fiete. 2017. “Training Recurrent Networks to Generate Hypotheses about How the Brain Solves Hard Navigation Problems.” In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 165 30:4529–38. Curran Associates, Inc.
- Kell, A. J. E., Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. 2018. “A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy.” *Neuron* 98 (3): 630–44.e16.
- 170 Kulkarni, Tejas D., William F. Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. “Deep Convolutional Inverse Graphics Network.” In *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, 28:2539–47. Curran Associates, Inc.
- Ladenbauer, Josef, Sam McKenzie, Daniel Fine English, Olivier Hagens, and Srdjan Ostojic. 2019. “Inferring and Validating Mechanistic Models of Neural Microcircuits Based on Spike-Train Data.” *Nature Communications* 10 (1): 4933.
- 175 LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–44.
- Lee, Hyodong, and James J. DiCarlo. 2019. “Topographic Deep Artificial Neural Networks (TDANNs) Predict Face Selectivity Topography in Primate Inferior Temporal (IT) Cortex.” *arXiv [q-bio.NC]*. arXiv. <http://arxiv.org/abs/1909.09847>.
- 180 Lerer, Adam, Sam Gross, and Rob Fergus. 2016. “Learning Physical Intuition of Block Towers by Example.” *arXiv [cs.AI]*. arXiv. <http://arxiv.org/abs/1603.01312>.
- Maheswaranathan, N., A. Williams, and M. Golub. 2019. “Universality and Individuality in Neural Dynamics across Large Populations of Recurrent Networks.” *Advances in Neural Information Processing Systems*. <http://papers.nips.cc/paper/9694-universality-and-individuality-in-neural-dynamics-across-large-populations-of-recurrent-networks>.
- 185 Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. “Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex.” *Nature* 503 (7474): 78–84.
- Marblestone, Adam H., Greg Wayne, and Konrad P. Kording. 2016. “Toward an Integration of Deep Learning and Neuroscience.” *Frontiers in Computational Neuroscience* 10 (September): 94.
- 190 Mastrogiuseppe, Francesca, and Srdjan Ostojic. 2018. “Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks.” *Neuron* 99 (3): 609–23.e29.
- Michaels, Jonathan A., Benjamin Dann, and Hansjörg Scherberger. 2016. “Neural Population Dynamics during Reaching Are Better Explained by a Dynamical System than Representational Tuning.” *PLoS Computational Biology* 12 (11): e1005175.
- 195 Michaels, Jonathan A., Stefan Schaffelhofer, Andres Agudelo-Toro, and Hansjörg Scherberger. 2019. “A Neural Network Model of Flexible Grasp Movement Generation.” <https://doi.org/10.1101/742189>.
- Nalisnick, Eric, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. 2018. “Do Deep Generative Models Know What They Don’t Know?” *arXiv [stat.ML]*. arXiv. <http://arxiv.org/abs/1810.09136>.
- 200 Nishimoto, Shinji, and Jack L. Gallant. 2011. “A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 31 (41): 14551–64.
- Remington, Evan D., Devika Narain, Eghbal A. Hosseini, and Mehrdad Jazayeri. 2018. “Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics.” *Neuron* 98 (5): 1005–19.e5.
- 205 Richards, Blake A., Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, et al. 2019. “A Deep Learning Framework for Neuroscience.” *Nature Neuroscience* 22 (11): 1761–70.
- Russo, Abigail A., Sean R. Bittner, Sean M. Perkins, Jeffrey S. Seely, Brian M. London, Antonio H. Lara, Andrew Miri, et al. 2018. “Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response.” *Neuron* 97 (4): 953–66.e8.
- Sarafyazd, Morteza, and Mehrdad Jazayeri. 2019. “Hierarchical Reasoning by Neural Circuits in the Frontal

- 210 Cortex." *Science* 364 (6441). <https://doi.org/10.1126/science.aav8911>.
- Schrimpf, Martin, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, et al. 2018. "Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like?" *bioRxiv*. <https://doi.org/10.1101/407007>.
- 215 Shepard, R. N., and J. Metzler. 1971. "Mental Rotation of Three-Dimensional Objects." *Science* 171 (3972): 701–3.
- Sohn, Hansem, Devika Narain, Nicolas Meirhaeghe, and Mehrdad Jazayeri. 2019. "Bayesian Computation through Cortical Latent Dynamics." *Neuron* 103 (5): 934–47.e5.
- Spearman, Charles. 1961. "The Proof and Measurement of Association between Two Things." <https://psycnet.apa.org/record/2006-10257-005>.
- 220 Storrs, Katherine R., and Nikolaus Kriegeskorte. 2019. "Deep Learning for Cognitive Neuroscience." *arXiv [q-bio.NC]*. arXiv. <http://arxiv.org/abs/1903.01458>.
- Sussillo, David, and Omri Barak. 2013. "Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks." *Neural Computation* 25 (3): 626–49.
- 225 Sussillo, David, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. 2015. "A Neural Network That Finds a Naturalistic Solution for the Production of Muscle Activity." *Nature Neuroscience* 18 (7): 1025–33.
- Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (6022): 1279–85.
- 230 Ullman, Tomer D., Elizabeth Spelke, Peter Battaglia, and Joshua B. Tenenbaum. 2017. "Mind Games: Game Engines as an Architecture for Intuitive Physics." *Trends in Cognitive Sciences* 21 (9): 649–65.
- Wang, Jing, Devika Narain, Eghbal A. Hosseini, and Mehrdad Jazayeri. 2018. "Flexible Timing by Temporal Scaling of Cortical Responses." *Nature Neuroscience* 21 (1): 102–10.
- 235 Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex." *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8619–24.
- Yang, Guangyu Robert, Michael W. Cole, and Kanaka Rajan. 2019. "How to Study the Neural Mechanisms of Multiple Tasks." *Current Opinion in Behavioral Sciences* 29 (October): 134–43.
- 240 Yang, Guangyu Robert, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. 2019. "Task Representations in Neural Networks Trained to Perform Many Cognitive Tasks." *Nature Neuroscience* 22 (2): 297–306.
- Zacks, Jeffrey M. 2008. "Neuroimaging Studies of Mental Rotation: A Meta-Analysis and Review." *Journal of Cognitive Neuroscience* 20 (1): 1–19.