A Domain-general Strategy for Hidden-state Inference in Humans and Noisy Neural Networks

Junseok K. Lee (jun.seok.lee@ens.psl.eu)
Laboratoire de Neurosciences Cognitives et Computationnelles,
Institut National de la Santé et de la Recherche Médicale,
Département d'Études Cognitives, École Normale Supérieure, Université PSL,
Paris 75005, France

Valentin Wyart (valentin.wyart@ens.psl.eu)
Laboratoire de Neurosciences Cognitives et Computationnelles,
Institut National de la Santé et de la Recherche Médicale,
Département d'Études Cognitives, École Normale Supérieure, Université PSL,
Paris 75005, France

Abstract:

Understanding the hidden (latent) states and structures that generate observations of the world is a fundamental aspect of cognition, wherein humans demonstrate exceptional proficiency in the ability to apply similar cognitive strategies across superficially dissimilar contexts sharing the same latent structure. While previous efforts to understand the computational bases of these cognitive strategies through cognitive modeling have largely focused on single contexts, here we take a novel approach which combines two tasks with the same reversal structure. Through cognitive modeling, we first show that humans use the same noisy hidden-state inference strategy across these superficially dissimilar tasks that require reversal learning. Then, using recurrent neural networks (RNNs) featuring either exact or noisy computations trained on the same tasks, we show that noisy RNNs – like humans – benefit from reusing the same latent representations for solving the two tasks. Together, our findings underscore the significance of computation noise in constraining the use of mental resources, shedding light on its potential functional role in cognition.

Keywords: human; behavior; hidden-state inference; learning; computational modeling; recurrent neural networks

Introduction

The capacity to use the same efficient cognitive strategies across diverse contexts is a defining characteristic of human cognition (Collins & Frank, 2013; Lake et al., 2017). Humans have been shown to spontaneously generate and use latent states and structures, which can then be used across superficially dissimilar contexts (Collins, Cavanagh & Frank, 2014). However, the study of these hidden-state inferences through the lens of computational model parameters have had challenges (Eckstein et al., 2022) and are usually confined to specific paradigms (e.g., reinforcement learning; Franklin & Frank, 2020). By combining two tasks with the same reversal structure, manifested as either rewards of a reinforcement learning task or colors in a perceptual categorization task, we show that humans use the same noisy hidden-state inference strategy across these superficially dissimilar contexts, shown through behavior and parameter fits of a cognitive computational model. Furthermore, the use of RNNs trained to perform both tasks offers evidence that this mechanism may be the natural result of a constrained alignment of the cross-context latent structure when task performance is optimized in the presence of computation noise.

Results

Task. Two reversal-learning tasks generated from the same latent structure from which stimulus values are sampled (Figure 1a) dictates reward outcomes for one option in the two-armed reward-guided learning (bandit) task (Figure 1b) while simultaneously assigning perceptual stimuli (i.e., colors) for a two-option perceptual categorization (fairy) task (Figure 1c). In the bandit task, participants learn to maximize their rewards by learning which of the two arms is more rewarding, while in the fairy task they infer from which bag an invisible fairy is currently drawing apples. For any participant, the trajectory of the latent state of a block and reversal points of one task correspond identically to that of a block of the other task (72 trials per block; 4 blocks per task). The tasks differ since their stimuli are visually dissimilar (i.e., numeric rewards and colors between green and red) and that bandit task stimuli are outcomes of actions while fairy task stimuli are uncontrollable observations. After the behavioral experiment, participants also responded to the 16-item International Cognitive Ability Resource (ICAR) test (Condon & Revelle, 2014).

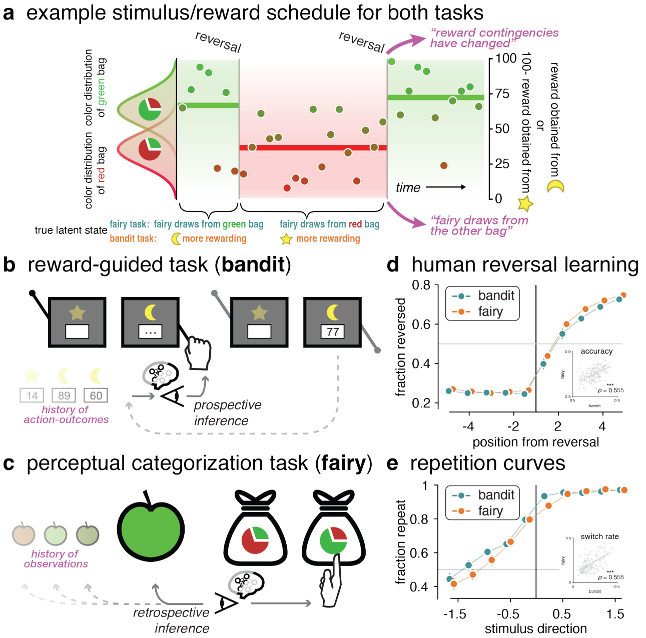

Figure 1: *Behavioral task and human performance*

Behavioral results. Participants ($N = 149$) achieved similar accuracy in both tasks (0.70 ± 0.01 in both tasks; $z = 1.24$, p $= .215$; across-task correlation $r = 0.56$, p $<$ $.001$; Figure 1d). They also showed similar signatures of reversal learning (Figure 1d) and sensitivity to the value of the stimulus such that they tended to repeat actions when the current stimulus (reward or color) was in agreement with their previous choice (Figure 1e). A subset ($N = 106$) of this original group also performed the same task two weeks later in a retest session.

Computational modeling results. The fitted parameters of a variant of the hazard-rated Bayesian inference model with noisy inference (Glaze et al. 2015; Weiss et al., 2021) showed that while the parameters responsible for stimulus processing and choice policy differed significantly between the two tasks, parameters responsible for inference did not. Furthermore, inference parameters, specifically the hazard rate and inference noise, showed interindividual stability across both tasks and sessions (Figure 2a). ICAR scores (here, used as a proxy measure of cognitive ability) solely predict the value of inference noise, negatively ($p < .001$ for all four regressions), but was not related to any other parameter of the model in both bandit and fairy tasks at test and retest. Structural equation modeling (Figure 2b) showed that cognitive ability may be causally related to inference noise in both tasks through a hierarchical structure (h1, $p > .989$) that assumes a general cognitive precision across tasks (Drugowitsch et al., 2016) but also shows context adaptation (see also Lee, Rouault & Wyart, 2023). This structural model was more plausible given the human data than alternative hypotheses assuming either context-agnostic or context-specific inference noise.

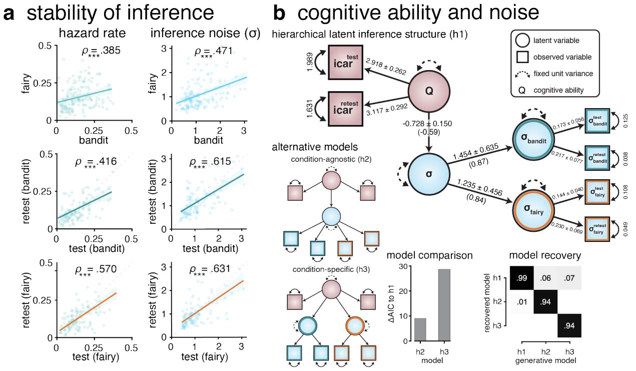

***Figure 2:** Stability of inference parameters and structure of the relation between cognitive ability and noise*

Recurrent Neural Network results. Recurrent neural networks (RNNs) are often employed to study learning and meta-learning across domains (Barak, 2017; Sheahan et al., 2021). However, interpreting derivative measures from networks and generating hypotheses from them may be challenging if they are not subject to biologically-plausible constraints (Pulvermüller et al., 2021; Schaeffer, Khona & Fiete, 2022). To investigate the role of noise in inference stemming from shared constrained resources, we trained recurrent neural networks with saturating units of varying hidden layer (HL) sizes, constrained by noisy computations to reproduce the presence of noise in human inferences (Findling and Wyart, 2021).

RNNs with hyperbolic tangent activation functions at the HL (configured as in Figure 3a) were trained to perform both tasks on newly generated blocks for each task, optimizing for task accuracy across the two tasks.

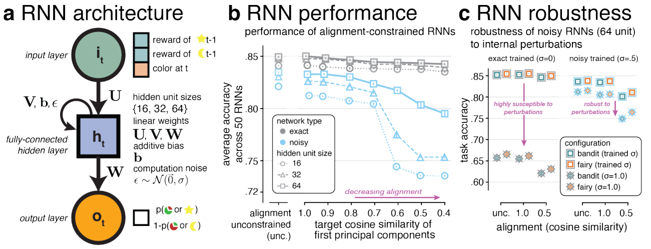

***Figure 3:** Effect of noise and alignment in RNNs*

Overall, exact RNNs outperformed noisy ones, where task accuracy declined as network size decreased (Figure 3b). To explore the mechanism underlying cross-context latent-state inference, we examined the presence of shared representations during RNN inference via the cosine similarity of the first principal components (PC1s) of HL activations in each task. Both noisy and exact RNNs showed high PC1 cosine similarity. To determine whether the use of shared latent representations for both tasks was critical to RNN inference, we trained RNNs to maximize task accuracy in an adversarial fashion by aiming for a target cosine similarity of PC1s lower than the one achieved when unconstrained. Exact RNNs were minimally affected by this additional constraint. However, noisy RNNs suffered substantially reduced performance as hidden layer alignment decreased, with pronounced interactions with lower HL sizes (Figure 3b).

Discussion

Our investigation found that humans may be using shared inferential mechanisms across superficially different tasks with matched latent structures, and that their cognitive ability is linked with the precision of these inferences. Likewise, we found that unconstrained artificial RNNs – whether precise or noisy – demonstrated proficient performance in the two tasks. However, under computation noise constraints, RNNs required alignment of hidden-unit activations to perform both tasks at near-optimal performance. This second result indicates a necessity for shared representations of inference to accomplish optimal performance across tasks in the presence of computation noise. Together, these findings suggest that understanding the human capacity for cross-context inference may be the result of optimization of objectives given the presence of biological noise.

Beyond biological plausibility, noisy RNNs have other advantages (Findling and Wyart, 2020). Noisy-trained RNNs were robust to internal perturbations (noise acting as a form of functional regularizer), whereas their exact counterparts significantly declined in accuracy (Figure 3c). Further research may uncover other advantages (Ma, Yan & Tang, 2023) such as resilience to external uncertainties, along with disadvantages due to excessive computation noise (Tran et al., 2020).

Acknowledgments

This work was supported by a starting grant from the European Research Council awarded to V.W. (ERC-StG759341), and by an institutional grant from the Agence Nationale de la Recherche awarded to the Département d'Etudes Cognitives (ANR-17-EURE-0017, EUR FrontCog).

References

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, *46*, 1-6.

Collins, A.G., & Frank, M.J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

Collins, A.G., Cavanagh, J.F., & Frank, M.J. (2014). Human EEG uncovers latent generalizable rule structure during learning. *Journal of Neuroscience*, *34*(13), 4677-4685.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64.

Drugowitsch, J., Wyart, V., Devauchelle, A.D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, *92*(6), 1398-1411.

Eckstein, M.K., Master, S.L., Xia, L., Dahl, R.E., Wilbrecht, L., & Collins, A.G. (2022). The interpretation of computational model parameters depends on the context. *Elife*, *11*, e75474.

Franklin, N.T., & Frank, M.J. (2020). Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS computational biology*, *16*(4), e1007720.

Findling, C., & Wyart, V. (2020). Computation noise promotes cognitive resilience to adverse conditions during decision-making. *BioRxiv*, 2020-06.

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124-132.

Glaze, C. M., Kable, J. W., & Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *Elife*, *4*, e08825.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., & Gershman, S.J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lee, J.K., Rouault, M., & Wyart, V. (2023). Adaptive tuning of human learning and choice variability to unexpected uncertainty. *Science Advances*, *9*(13), eadd0501.

Ma, G., Yan, R., & Tang, H. (2023). Exploiting noise as a resource for computation and learning in spiking neural networks. *Patterns*, *4*(10).

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., & Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, *22*(8), 488-502.

Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, *35*, 16052-16067.

Sheahan, H., Luyckx, F., Nelli, S., Teupe, C., & Summerfield, C. (2021). Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, *109*(7), 1214-1226.

Tran, T. T., Rolle, C. E., Gazzaley, A., & Voytek, B. (2020). Linked sources of neural noise contribute to age-related cognitive decline. *Journal of cognitive neuroscience*, *32*(9), 1813-1822.

Weiss, A., Chambon, V., Lee, J. K., Drugowitsch, J., & Wyart, V. (2021). Interacting with volatile environments stabilizes hidden-state inference and its brain signatures. *Nature communications*, *12*(1), 2228.