Decoding Latent Attention Across Cognitive Models

Christina Maher
Friedman Brain Institute
Icahn School of Medicine at Mount Sinai
New York, NY 10029
christina.maher@icahn.mssm.edu

Ignacio Saez
Depts. Neuroscience, Neurology, Neurosurgery
Icahn School of Medicine at Mount Sinai
New York, NY 10029
Ignacio.saez@mssm.edu

Angela Radulescu
Department of Psychiatry
Icahn School of Medicine at Mount Sinai
New York, NY 10029
angela.radulescu@mssm.edu

Abstract

Selective attention plays a critical role in representation learning [1], with two competing computational cognitive models proposing distinct mechanisms by which attention arises. Feature-based reinforcement learning (FRL) posits selective attention arises through retrospective value learning, while serial hypothesis testing (SHT) suggests selective attention arises via prospective hypothesis testing. Here, we apply LaseNet [2], a novel neural network method for directly inferring latent variables from cognitive models with both synthetic and human data, to decode trial-by-trial attention as agents learn in a multidimensional environment. Networks trained on data generated from SHT models outperformed networks trained on data from FRL models in predicting attention in a labeled human dataset. SHT networks also showed limited generalizability to unseen data across model classes, reflecting distinct mechanisms for attention allocation under FRL and SHT models. This work utilizes a cutting-edge approach to infer attention dynamics from cognitive models, significantly enhancing their evaluation and providing deeper insights into the attention mechanisms that drive human representation learning. By leveraging this method, we can uncover latent attention processes underlying human representation learning, ultimately informing model-based neural analyses.

Keywords: Representation learning, multi-dimensional reinforcement learning, selective attention, serial hypothesis testing, LaseNet

1 Introduction

Previous research highlights the role of selective attention in state representation learning [1]. Two main classes of computational cognitive models, each supported by choice data and neural evidence, propose competing mechanisms for implementing selective attention in multidimensional environments: feature-based reinforcement learning (FRL) models suggest that selective attention emerges retrospectively through value learning of stimulus features [3]. In contrast, serial hypothesis testing (SHT) models propose a prospective process, where selective attention is dynamically allocated by iteratively testing hypotheses about which features of the task are relevant [4], [5]. Hybrid models that incorporate elements of both FRL and SHT have also been proposed [6].

Traditional model evaluation methods based on maximum likelihood estimation (MLE) are limited to models for which analytically tractable likelihoods can be derived and computed [7]. SHT models, and sampling models more generally, are challenging analytically, and typically intractable due to combinatorial explosions. This has constrained cognitive models of attention to be biased towards FRL models. However, mounting evidence suggests that hypothesis sampling plays a key role in attention allocation [4], [5], [6]. Here, we use LaseNet, a novel method for direct inference of latent cognitive variables [2], to directly decode attention generated from different models in the FRL and SHT family. Our goal is to arbitrate between these model classes, and to generate precise predictions of attention allocation for model-based neural analysis [8].

We hypothesized that: (1) networks trained on SHT models would label human participants' trial-by-trial hypotheses more accurately than those trained on FRL models, and (2) networks trained using a given class of models would not generalize well to data generated by models from a different model class. This would indicate that the two model classes capture distinct cognitive processes which the network is able to discern. To test these hypotheses, we trained LaseNet neural estimators to infer trial-by-trial attention allocation (i.e., their hypothesis about the most relevant feature) using synthetic data generated from each cognitive model. We evaluated the networks on held-out synthetic test data, and on a self-labeled human-generated dataset. All networks performed above chance, with SHT models outperforming FRL models in labeling the human dataset. Additionally, networks did not generalize well across model classes. Findings highlight the distinct mechanisms underlying representation learning proposed by FRL and SHT and provide a foundation for adjudicating between their competing theories.

2 Task

To train neural networks, we used a multidimensional RL task adapted from prior work [3], [9]. In the human experiment, participants (N=21 neurosurgical patients) completed six 18-trial games in which they made repeated choices between three stimuli varying in shape (square, oval, circle) and color (orange, yellow, blue) (**Fig 1A**). Each game had one relevant dimension (shape or color) and a target feature (e.g., square). After each choice, participants were rewarded with 80% probability if they selected the stimulus containing the target feature. To maximize reward, participants had to learn the target feature via trial-and-error. Changes in the relevant dimension were explicitly signaled between games. And participants were aware of the generative structure of the task (i.e. one target feature being relevant, and the exact reward probabilities).

3 LaseNet Estimators

We decoded trial-by-trial attention in the multidimensional RL task using five cognitive models. We used LaseNet [2], a novel technique based on neural Bayes estimation which directly maps choice data to latent variable space by using recurrent neural networks trained on synthetic data generated by a cognitive model. Unlike traditional maximum likelihood estimation, which requires predefined parameter space comparisons, LaseNet directly infers latent variables from behavior, enabling the use of models with analytically intractable or computationally intensive likelihoods. During the training phase (**Fig. 1B**, adapted from [2]), we create a synthetic dataset by simulating the desired cognitive model. LaseNet is trained using model-simulated observable data Y as input and a series of model-derived latent variables Z as output. During the inference phase (**Fig. 1C**), the trained LaseNet takes the observable experimental data as input to infer a sequence of unobservable latent variables. In this project, our goal was to infer participants' attention allocation by decoding their trial-by-trial hypotheses H about the target feature. For each LaseNet estimator, we simulated 20000 (Z, Y) pairs with 720 trials in each pair as training data. In effect, this means that each individual

parameter setting ("agent") generated 40 games, each representing an instance of dynamic attention allocation with that parameter setting. While FRL models would produce fixed attention trajectories with the same parameter setting, SHT models change attention stochastically, so each game represents a possible trajectory through hypothesis space. We simulated an additional unseen 20 (Z, Y) pairs with 720 trials each as testing data. Each pair was generated with uniform priors on model parameters θ. Although the hypotheses participants are testing are latent in an experimental context, we collected one exploratory dataset (6 games, 18 trials/game, 108 trials in total) in which the participant was instructed to self-label their hypothesis on each trial. Although we have access to these self-reported hypotheses, we do not have access to the generative model that produced their choices. This discrepancy allows us to assess the ability of LaseNet Estimator trained on different cognitive models to infer the participants' latent hypotheses.

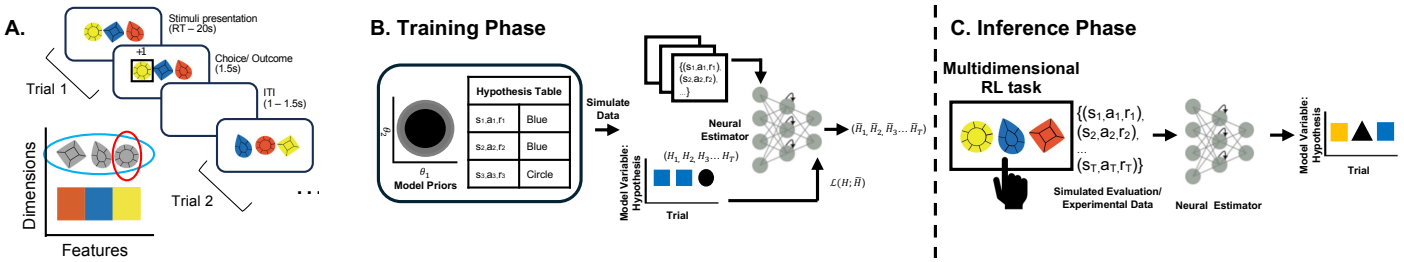

Fig 1. LaseNet method applied to multidimensional RL. A. Multidimensional RL task. **B.** Network is trained to predict latent variables from a cognitive model (i.e., target feature hypothesis) using simulated data. Input includes trial-wise observable data (stimuli, actions, rewards). **C.** Trained networks predict latent variable for experimental data with unknown ground truth. Schematic adapted from [2].

4 Cognitive Models

Two cognitive model classes have emerged to capture attention during state representation learning: FRL [3], [9] and SHT [4], [5], [6]. Models within these respective classes seek to explain how attentional mechanisms support efficient state representation learning by prioritizing relevant information. FRL models propose agents learn to assign values to individual features of stimuli, which are then integrated to guide their choices. Attention is dynamically allocated to specific features based on their perceived value for maximizing future rewards. FRL is an extension of traditional RL, where the focus shifts from learning the value of entire stimuli to learning the value of discrete features within those stimuli. In this context, value learning over time drives attentional allocation towards the features that most reliably predict reward. We trained LaseNet Estimators to decode latent hypotheses using two FRL models: FRL and FRL with decay [9]. The FRL model maintains that participants learn and update values for each feature in the environment. To account for human's limited working memory capacity, the FRL with decay model decays the value of nonchosen features. At each timestep, the hypothesis H_t is taken to be the feature with the highest value.

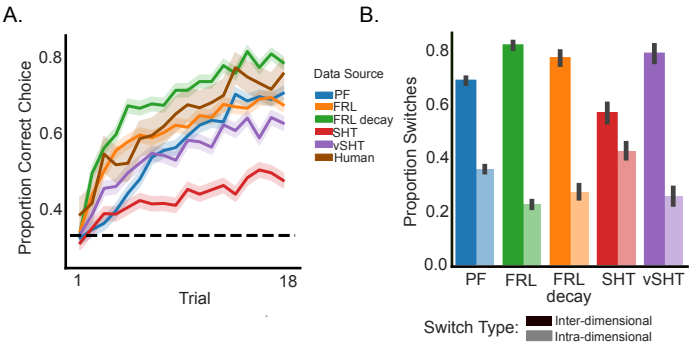

Fig 2. Cognitive Model Validation. A. Proportion of correct choice increases across trials in simulated (N=20; 40 games, 18 trial/game) and human (N=21; 6 games, 18 trials/game; shading = SEM; dashed line = chance) data. **B.** Proportion of inter- vs intra-dimensional hypothesis switches by model (N=20 simulated agents, 40 games, 18 trials/game; error bars = SEM).

In contrast, SHT models suggest that efficient state representations arise through the evaluation of competing hypotheses regarding the most relevant environmental features. These models can be thought of as a tractable and computationally efficient approximations of full Bayesian inference. According to this framework, representation learning involves maintaining a single hypothesis H_t about the relevant feature and iteratively updating it as new information is received. Attention is directed toward the feature that the agent hypothesizes is most rewarding. To decode latent hypotheses, we trained LaseNet Estimators using three variants of hypothesis testing models: Serial Hypothesis Testing (SHT) [5], Value-based Serial Hypothesis Testing (vSHT) [6], and a Memory-Augmented Particle Filter (PF) [4].

5 Results

Before training LaseNet, we validated the cognitive models by comparing the performance of synthetic test data to human data on the same task. All models successfully learned the task (**Fig. 2A**) and exhibited attention-switching behavior indicative of adaptive learning, with more inter-dimensional switching (**Fig. 2B**).

Next, we trained five separate LaseNet Estimators for each of our cognitive models. To evaluate training, we assessed each network's ability to label hypotheses using a held-out test set (N = 20 agents; **Fig. 3A**). All networks performed above chance (**Fig 3B**). Then, we evaluated each network's ability to label trial-by-trial hypotheses in this self-labeled human dataset, in which the generative cognitive model is unknown. As we hypothesized, the SHT models performed better than the FRL models, indicating that the behavior in this dataset was more accurately captured by this model class (**Fig 3C**).

Finally, we assessed the generalizability of the PF- and vSHT-trained networks to determine whether they are more effective at identifying hypothesis testing behavior specific to their own generative models. To do this, we evaluated these networks' ability to label hypotheses using the synthetic test datasets generated by each cognitive model (**Fig. 4**). As hypothesized, model type had a significant effect on the network's labeling accuracy (PF: $F=7.2$, $p<0.0001$, **Fig 4A**; vSHT: $F=17.2$, $p<0.0001$, **Fig 4B**). Both PF and vSHT networks performed best when labeling test data generated by the model it was trained on, with stronger performance within its model class and reduced accuracy when generalizing to FRL models. These results suggest that the networks capture model-specific cognitive mechanisms that are not generalizable across model classes.

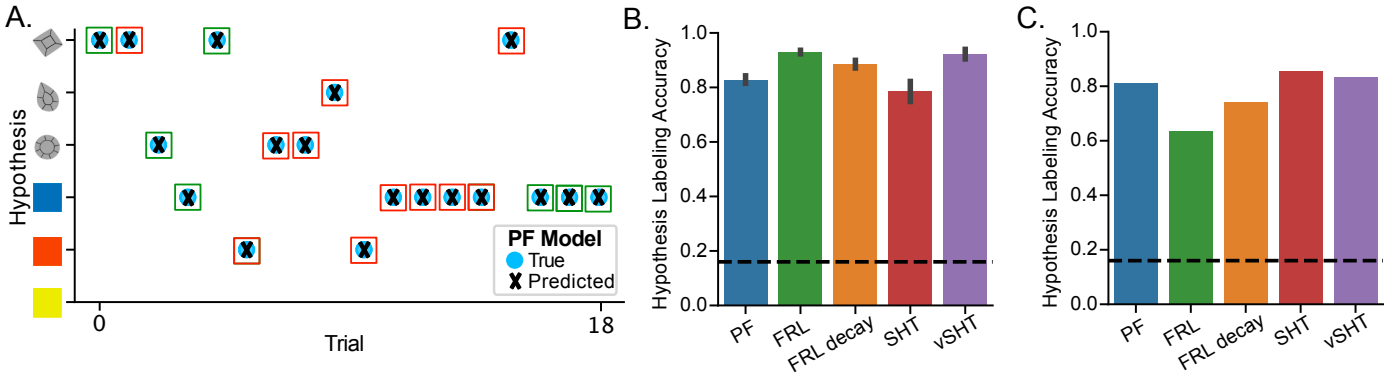

Fig 3. Evaluation of LaseNet Estimators on simulated and human data. A. PF-trained network's hypothesis labeling accuracy for one example game in which it correctly labeled every trial (blue circle = true hypothesis, black cross = predicted hypothesis, green = reward, red = no reward; target feature = yellow). **B.** Hypothesis labeling accuracy (represented by the proportion of correctly labeled trials, error bars = SEM) for LaseNet Estimators evaluated on synthetic test data (N = 20 agents; 40 games, 18 trials per game) for each cognitive model. All networks perform well above chance (dashed line). **C.** Hypothesis labeling accuracy (represented by the proportion of correctly labeled trials) for LaseNet Estimators evaluated on self-labeled human dataset (N = 1 participant; 6 games, 18 trials per game). All models perform significantly above chance (dashed line), however the hypothesis testing models (PF, SHT, vSHT) outperform the FRL models (FRL, FRL decay).

6 Discussion

We trained neural networks using LaseNet to infer latent hypothesis testing under two distinct cognitive model classes, FRL and SHT, which capture attentional mechanisms underlying state representation learning. This innovative approach to cognitive dynamics inference circumvents the limitations of traditional model fitting and comparison methods, enabling us to infer latent cognitive variables with high predictive accuracy, and allowing for direct comparison of inferences across different model assumptions.

Our results demonstrated that all networks performed well above chance in inferring hypotheses on simulated data (**Fig. 3B**). As hypothesized, the SHT models outperformed the FRL models in labeling human hypotheses from a self-labeled dataset with known ground truth (**Fig. 3C**). These findings suggest that the SHT models are more effective than FRL in capturing the cognitive processes underlying human state representation learning. Furthermore, the networks demonstrated low out-of-class generalizability (**Fig. 4**), which underscores their ability to isolate behavioral dynamics specific to each cognitive model.

Several limitations warrant consideration. One potential limitation is the use of a uniform prior during the training phase for all models, which may not accurately reflect the true parameter ranges observed in human data. This discrepancy could explain the relatively lower performance of the SHT and vSHT models (**Fig. 2A**). Although the uniform prior avoided potential training biases, incorporating more accurate priors, derived from previous studies, could help constrain the parameter space more effectively and improve the network's ability to capture realistic patterns in the data. Additionally, performing hyperparameter tuning on unseen simulated data prior to training could further optimize model performance.

An interesting aspect of our results lies within the SHT model class, which performed better at labeling human hypotheses (**Fig. 3C**) than FRL. Within this class, the SHT, vSHT, and PF models make different assumptions about how the proposal distribution is maintained. PF model captures state inference while accounting for working memory constraints, potentially making it a more biologically plausible model for human representation learning. Although we lack ground truth for participant behavior in experimental data, these competing models can be tested against neural data to identify which model's inferences is most reflective of observed neural activity. We are poised to investigate these inferences in our cohort of 21 neurosurgical participants who completed the task and have corresponding cortical/subcortical local field potential recordings. These data will help us assess the assumptions of different models and identify which most closely aligns with neuronal activity, shedding light on the biological basis of attention in state representation learning.

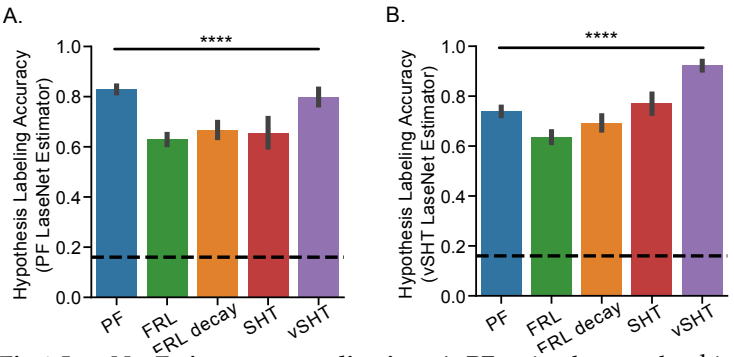

Fig 4. LaseNet Estimator generalization. A. PF-trained network achieves the highest hypothesis labeling accuracy on PF model data and performs better overall on SHT models compared to FRL models (N=20 agents, 40 games, 18 trials/game; error bars = SEM). **B.** vSHT-trained network achieves the highest hypothesis labeling accuracy on vSHT model data and performs better overall on SHT models compared to FRL models (N=20 agents, 40 games, 18 trials/game; error bars = SEM).

6 References

[1] A. Radulescu, Y. Niv, and I. Ballard, "Holistic Reinforcement Learning: The Role of Structure and Attention," *Trends Cogn. Sci.*, vol. 23, no. 4, pp. 278–292, Apr. 2019, doi: 10.1016/j.tics.2019.01.010.

[2] T.-F. Pan, J.-J. Li, B. Thompson, and A. Collins, "Latent Variable Sequence Identification for Cognitive Models with Neural Network Estimators," Dec. 19, 2024, *arXiv*: arXiv:2406.14742. doi: 10.48550/arXiv.2406.14742.

[3] Y. C. Leong, A. Radulescu, R. Daniel, V. DeWoskin, and Y. Niv, "Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments," *Neuron*, vol. 93, no. 2, pp. 451–463, Jan. 2017, doi: 10.1016/j.neuron.2016.12.040.

[4] A. Radulescu, Y. Niv, and N. Daw, "A particle filtering account of selective attention during learning," in *2019 Conference on Cognitive Computational Neuroscience*, Berlin, Germany: Cognitive Computational Neuroscience, 2019. doi: 10.32470/CCN.2019.1338-0.

[5] R. C. Wilson and Y. Niv, "Inferring Relevance in a Changing World," *Front. Hum. Neurosci.*, vol. 5, 2012, doi: 10.3389/fnhum.2011.00189.

[6] M. Song, P. A. Baah, M. B. Cai, and Y. Niv, "Humans combine value learning and hypothesis testing strategically in multi-dimensional probabilistic reward learning," *PLOS Comput. Biol.*, vol. 18, no. 11, p. e1010699, Nov. 2022, doi: 10.1371/journal.pcbi.1010699.

[7] B. Van Opheusden, L. Acerbi, and W. J. Ma, "Unbiased and efficient log-likelihood estimation with inverse binomial sampling," *PLOS Comput. Biol.*, vol. 16, no. 12, p. e1008483, Dec. 2020, doi: 10.1371/journal.pcbi.1008483.

[8] C. Maher, S. Qasim, L. N. Martinez, I. Saez, and A. Radulescu, "Intracranial recordings reveal neural encoding of attention-modulated reinforcement learning in humans," in *Computational Cognitive Neuroscience*, 2024. https://2024.ccneuro.org/pdf/548_Paper_authored_CCN2024_Maheretal.pdf.

[9] Y. Niv *et al.*, "Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms," *J. Neurosci.*, vol. 35, no. 21, pp. 8145–8157, May 2015, doi: 10.1523/JNEUROSCI.2978-14.2015.