COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e70015

ISSN: 1551-6709 online
DOI: 10.1111/cogs.70015

Inverting Cognitive Models With Neural Networks to Infer Preferences From Fixations

Evan M. Russek,[a] Frederick Callaway,[b,c] Thomas L. Griffiths[a,d]

[a]*Department of Computer Science, Princeton University*
[b]*Department of Psychology, New York University*
[c]*Department of Psychology, Harvard University*
[d]*Department of Psychology, Princeton University*

Received 8 January 2024; received in revised form 28 August 2024; accepted 21 October 2024

Abstract

Inferring an individual's preferences from their observable behavior is a key step in the development of assistive decision-making technology. Although machine learning models such as neural networks could in principle be deployed toward this inference, a large amount of data is required to train such models. Here, we present an approach in which a cognitive model generates simulated data to augment limited human data. Using these data, we train a neural network to invert the model, making it possible to infer preferences from behavior. We show how this approach can be used to infer the value that people assign to food items from their eye movements when choosing between those items. We demonstrate first that neural networks can infer the latent preferences used by the model to generate simulated fixations, and second that simulated data can be beneficial in pretraining a network for predicting human-reported preferences from real fixations. Compared to inferring preferences from choice alone, this approach confers a slight improvement in predicting preferences and also allows prediction to take place prior to the choice being made. Overall, our results suggest that using a combination of neural networks and model-simulated training data is a promising approach for developing technology that infers human preferences.

Keywords: Fixation; Cognitive models; Neural networks; Inverse reinforcement learning

Correspondence should be sent to Evan M. Russek, Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA. E-mail: erussek@princeton.edu

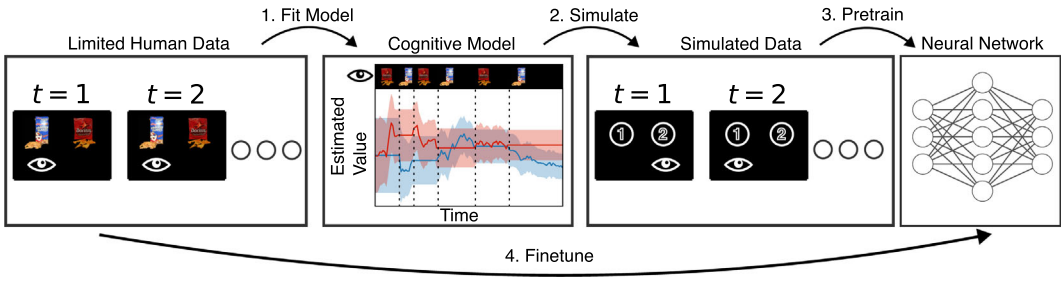

Fig. 1. We propose an approach to developing AI systems that estimate the latent variables that underlie human behavior. The approach is aimed at satisfying neural networks' need for massive data by using simulated data generated from a cognitive model. First, limited human data is used to fit a cognitive model. In the example we tackle in this paper, human data consists of eye fixations between food items along with self-reported preferences over those items. The cognitive model is a resource-rational model specifying how individuals select fixations. Each fixation is modeled as an information-gathering action which decreases uncertainty around the utility of an item. Second, the cognitive model is then used to simulate massive amounts of simulated data. Third, this simulated data is used to pretrain neural networks. Finally, the neural networks are additionally fine-tuned with limited human data.

Key to building systems that help people make better choices is inferring what people want from their behavior (Hadfield-Menell, Dragan, Abbeel, & Russell, 2016). How can this inference take place? Cognitive models, which specify how latent preferences generate behavior, could in principle be applied to this problem. By using Bayesian inference to invert such a model, we can infer preferences from behavior. However, cognitive models often fail to capture idiosyncratic relationships between preferences and behavior, and inverting such models is computationally burdensome. In contrast, machine learning models such as neural networks offer a way to make inference computationally feasible and have greater flexibility to capture arbitrary relationships. However, training such models requires vast amounts of behavioral data.

In this work, we propose and test a new solution to the problem of inferring preferences from behavior, combining the strengths of cognitive models and neural networks. Our approach is to satisfy the need for massive data to train neural networks by augmenting limited available real human data with simulated data from a cognitive model (Fig. 1). We apply this approach to the problem of inferring human preferences over food items from visual fixations between those items made during the decision-making process. Our results demonstrate that neural networks are able to learn, from simulated data, to invert a computationally intensive cognitive model for how individuals decide where to fixate while making a decision given their preferences over items. Additionally, pretraining a network with simulated data and fine-tuning with limited human data allows prediction of people's self-reported preferences from their fixations. This demonstrates a new approach for how cognitive models can be used to address key limitations of deploying neural networks in human-interaction systems.

1. Background

Our approach draws upon ideas from machine learning and cognitive modeling. In this section, we briefly review these ideas.

1.1. Inverse reinforcement learning

In machine learning, the problem of inferring another agent's preferences has been cast as inverse reinforcement learning (IRL; Ng & Russell, 2000). IRL specifies a generative model whereby agents have latent preferences (formalized as a utility function over task states and/or actions) and make decisions that maximize those preferences. This generative model, relating preferences to behavior, is inverted to predict the maximum a posteriori (MAP) preferences that generated the observed behavior. This general framework of inferring preferences by inverting a decision model has also formed the basis of cognitive models for how individuals make inferences about others preferences based on their behavior (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jern, Lucas, & Kemp, 2017; Jara-Ettinger, 2019; Lucas et al., 2014). Cognitive science has also recently provided more sophisticated models of how humans make decisions, which can provide more accurate models relating preferences to actions to guide inference (Ho & Griffiths, 2022), and can expand the observables over which inference can occur to data beyond choices (e.g., response times; Gates, Callaway, Ho, & Griffiths, 2021).

Although IRL defines how preference inference can occur in principle, its practical use has been limited by the computational challenge of inverting decision models. Finding the MAP preferences typically involves searching over, and computing the likelihood of, candidate utility functions. For many cognitive process models, computing this likelihood for a single utility function can be quite computationally intensive. This makes a full search process too computationally expensive to be deployed in real-time inference settings. As a step toward making inference faster, recent work has shown that it is possible to implement IRL in neural networks, for which inference is fast (Rabinowitz et al., 2018). However, this approach requires large amounts of labeled training data, which is often unavailable for real-life applications. Here, we test whether the use of simulated data can alleviate this need for real human data.

1.2. Modeling the relationship between fixations and choice

Rather than trying to infer preferences purely from choices, we consider the problem of predicting preferences from visual fixations. When individuals make a choice between items, they tend to move their gaze between potential items in a stereotypic manner. This process has been studied experimentally in tasks where a participant is presented with a screen displaying snack items, and is required to select which of them they would prefer to eat at the end of the experiment (Krajbich, Armel, & Rangel, 2010; Krajbich & Rangel, 2011). Recent work suggests that when making such decisions, people fixate on the different options in a way that depends on independently provided ratings of how much they like those items (Callaway, Rangel, & Griffiths, 2021; Gluth, Kern, Kortmann, & Vitali, 2020; Jang, Sharma, & Drugowitsch, 2021). These relationships in principle make it possible to predict individuals' utility over items from their fixations. Prior studies have found that it is possible to use the total as

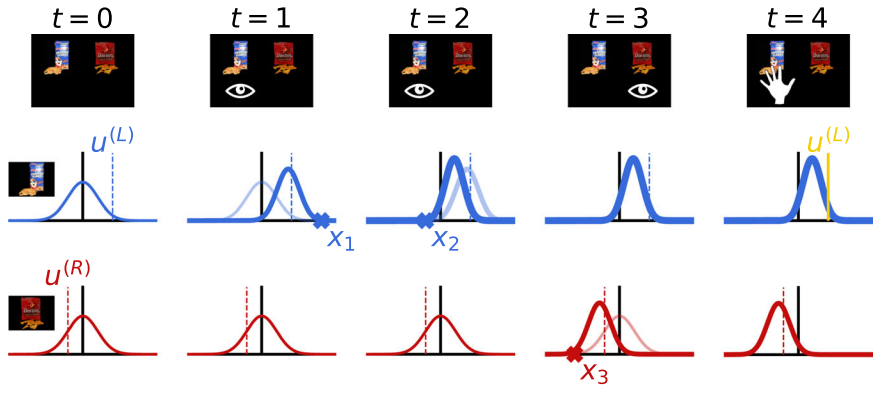

Fig. 2. We apply the proposed method to a rational model of attention allocation in preferential choice (Callaway et al., 2021). The top row shows the experimental display, with the currently fixated item (measured with an eye-tracker) denoted by the eye symbol. The bottom two rows depict the internal state of the cognitive model. The model maintains Gaussian beliefs about the subjective value of each item. The true subjective values (dashed lines) are sampled from a Gaussian; this is captured in the decision maker's initial belief state (first column). At every time step, t, the model "fixates on" one of the items and receives a noisy sample about the true value of that item (x_t marks). It then updates the belief about the value of the fixated item using Bayesian updating (shift from light to dark curve). The beliefs for the unfixated item are not updated. The process repeats at each time step until the model decides to make a choice, selecting the item with maximal posterior mean (hand icon). Decisions about when to make a choice and which item to fixate on at each moment are made by a policy that optimizes the expected value of the chosen item (yellow line) minus a cost proportional to the amount of time spent sampling.

well as proportion of time individuals spend fixating on different items to predict, to some extent, individuals' preferences for those items (Goyal, Miyapuram, & Lahiri, 2015; Glaholt, Wu, & Reingold, 2009).

We aim to assess how a cognitive model of how individuals select fixations can be used to improve this inference. Callaway et al. (2021) presented a resource-rational model for how individuals select both where to fixate at any point in time, and when to stop fixating and make a choice in such tasks. According to the model, eye movements reflect optimally selected information-gathering computations that improve the participant's beliefs about the utilities of different snack items (Fig. 2). These computations can lead to a better ultimate decision; however, they also incur a cognitive cost. By formalizing this process as a sequential decision problem (specifically, a meta-level Markov decision process), the optimal fixation policy can be identified. It was found that the sequences of fixations made by the optimal policy closely corresponded to participant's observed fixation behavior. We provide a brief summary of the model in the supplement and refer the reader to Callaway et al. (2021) for further details.

2. Inverting cognitive models using neural networks

Directly inverting cognitive models, such as the model of fixations in choice proposed by Callaway et al. (2021), is computationally infeasible. We thus propose a new approach:

using simulated data from this cognitive model to train neural networks to infer an individual's preferences given their fixations. Our task involves mapping a sequence (fixations) to a scalar (utility), so we use three types of neural networks that have been shown to be successful in handling sequence data: long-short-term-memory (LSTM) neural networks (Hochreiter & Schmidhuber, 1997), gated recurrent unit (GRU) neural networks (Cho, van Merrienboer, Bahdanau, & Bengio, 2014), and Transformers (Vaswani et al., 2017). LSTMs and GRUs are both variants of recurrent neural networks (RNNs), which map inputs to a hidden-unit representation, which in turn is mapped both to an output estimate and also provided back into itself as additional input for the next time-step. Through learning, the hidden unit representation comes to summarize the relevant input history of the sequence up to that time-point, thus enabling the output to be based not only on current input but also on a history of input. Whereas for original RNNs (Rumelhart, Smolensky, McClelland, & Hinton, 1986), hidden units passed their activation to the next time-step through a set of learned weights, GRUs and LSTMs can learn gates and other structures that control how much past information in a sequence should be passed along to future time-steps in a contextual manner.

Unlike RNNs, which update an internal representation one step at a time, Transformers process the entire sequence at once, mapping it to an internal representation through a "self-attention" mechanism (at test time, the model makes predictions sequentially given the observations before the current predicted element). In self-attention, each item in the input sequence is represented as a weighted sum of values derived from the entire sequence. These values are learned, allowing the weighted sum to capture relevant information from both the current item and other items in the sequence.

The attention weights are determined by an attention score, which measures the relevance of each item to the current one. This score is computed based on the similarity between the "query" vector of the current item and the "key" vector of the other items. Both key and query vectors are learned for each item to facilitate the task-specific relationships between items. Thus, each item's output representation is the sum of all items' value vectors, weighted by the attention scores between the current item's query vector and the other items' key vectors. This mechanism enables the model to incorporate relevant information from the entire sequence when representing each item.

We compare these sequence-based models to models using hand-defined sequential information (e.g., the sum of fixations up to a given time-point) to estimate utilities. For these comparisons, we use multi-layer perceptrons (MLPs). MLPs are neural networks that map features at a single time-point to a learned internal representation, which is then used to estimate the utility of each item.

Our approach builds on work in cognitive science and machine learning that has combined neural networks with simulated data to either invert complex generative models or to predict human choices. For fitting cognitive models to behavior, recent work has used neural networks to approximate likelihood functions that might otherwise be intractable (Fengler, Govindarajan, Chen, & Frank, 2021). Closer to our application here is work that has trained neural networks to directly estimate mean parameters or sample from posterior distributions of complex models, by training networks with simulated data labeled with corresponding parameters (Gonçalves et al., 2020; Ger, Nachmani, Wolf, & Shahar, 2023; Papamakarios

& Murray, 2016; Radev, Mertens, Voss, Ardizzone, & Kothe, 2022; Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020). Neural networks used to predict human decisions have been pretrained with simulated data from cognitive models to make up for limited real human data (Bourgin, Peterson, Reichman, Griffiths, & Russell, 2019). Finally, neural networks trained to predict human choices have in turn been used to improve cognitive models through a process referred to as Scientific Regret Minimization (Agrawal, Peterson, & Griffiths, 2020; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021; Kuperwajs, Schütt, & Ma, 2023).

We turn this approach toward the problem of estimating human preferences from eye fixations, training neural networks on simulated fixation and choice data from the model presented in Callaway et al. (2021). We first test whether we can simply invert the model; that is, we provide neural networks with a sequence of simulated fixations followed by a choice and test whether they output correct utilities over the three items. Following this, we validate the approach using real human data on a trinary choice task, reported in Krajbich and Rangel (2011). We determine whether neural networks can predict people's reported utilities given their fixations and choices, how this compares to prediction using choice alone, and also whether simulated data complements using human data alone in training models on this task.

3. Methods

3.1. Human data

Human data consisted of 2966 trials reported in Krajbich and Rangel (2011) in which participants made choices over three food items after having the opportunity to engage in a sequence of fixations between them. Fixations, f_{jt_i}, reflect the item most fixated on in a .1 second bin. Prior to all choices, participants provided liking ratings (utilities) over the full set of items.

3.2. Simulated data

Simulated data was generated using the model described in Callaway et al. (2021) (Fig. 2). To simulate a single trial, j, a utility, u_{js}, was drawn for each snack item, s, from $P(u)$, which was defined by fitting a Gaussian distribution to the full set of item ratings from Krajbich and Rangel (2011). Given such "true" utilities over items, the model generates a sequence of fixations, $f_{jt_i<T}$, over by items, followed by a choice, c_{jt_i}, $x_j = (f_{jt_1}, f_{jt_2}, \ldots, c_{jt_T})$. At a high level, each simulated fixation on item s collects a sample from a distribution of item utilities centered on u_{js}, with Gaussian noise. This sample is used to increase the accuracy of an estimate of that item's utility. Optimal fixations reflect the information gathering actions that balance the benefit of making a choice with a more accurate utility estimates with the cost of spending additional time. A detailed description of the generative model is presented in Supplementary section "Optimal fixation model." Parameters of the model are reported in Supplementary section "Parameters of optimal fixation model." Using this model, we simulated 1.5 million trials.

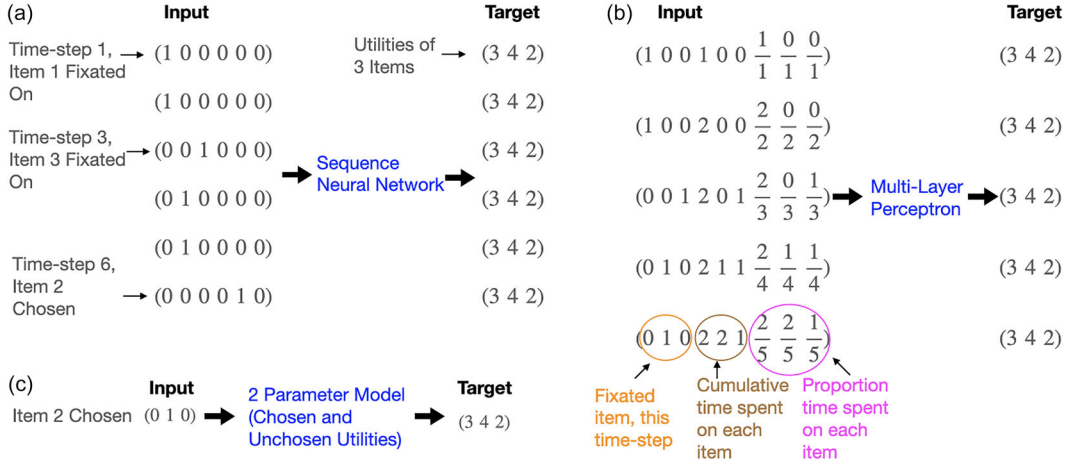

Fig. 3. Input and target data representation for an example trial. In this example trial, the participant fixated for two time-steps on item 1, one time-step on item 3, two time-steps on item 2, and then chose item 2. (a) Representation of example trial's input sequence and target sequence for training sequence-based models (e.g., Transformers). Note that to predict each item's utilities at time-point i, sequence-based models can make use of all time-points in input sequence up to and including time-point i. (b) Representation of example trial's input and target sequence for nonsequential control model with hand-designed features. Input representation of a time-point includes hand-designed features pertaining to cumulative information about fixations up to and including that time-point. Unlike sequence-based models, here multi-layer perceptrons have to map a single-time point's input representation to a prediction of each item's utilities. Note that this model uses only fixation, but not choice information. (c) Example trial's input and target representation for model trained on choice alone. This model just learns two parameters: one for the utility of the chosen item and one for the utility of the two unchosen items. Note that although in this case two unchosen items have different utility values, the choice-only model will assign the same prediction to these two items.

3.3. Input and target data representation

For each trial, j, consisting of T time-points, we represented that trial's sequence of fixations followed by a choice as a length-T sequence of 6-length vectors, x_{ji}, for each time-point, $i = 1 : T$ (Fig. 3a, "Input"). For each time-point, $i < T$, the first three elements of x_{ji} designated which of the three food items was fixated on at that time-point. The last three elements, which were active only for the final time-point, T, designated which of the three items was chosen on that time-point. Sequence-based models were trained to make a prediction of each of the three item's utilities at each time-point, i, in the sequence, using all input data up to time-point i. The target sequence thus consisted of a length-3 vector where each element, $j = 1 : 3$ contained the true utility of item j, u_j, repeated for each time-point in the sequence (Fig. 3a, "Target").

We compared models trained on both fixations and choice to a model trained on choice alone. For the model trained on choice alone, we trained a model that simply estimated two parameters reflecting the respective utilities of the chosen item and nonchosen items (Fig. 3c). We also defined a set of control models based on features that previous work has identified

as predictive of preferences: the cumulative total and proportion fixation time on each item (Goyal et al., 2015; Glaholt et al., 2009). For each time-point, we defined a length-9 vector, with three values indicating the current fixated item (ID), three indicating the total fixation time on each item (Sum), and three indicating the proportion fixation time on each item (Prop; Fig. 3b). We then trained MLPs to map these features at each time-point to utility estimates for each item. Performance of these models is compared in Fig. S1.

3.4. Training procedure and hyperparameter selection

Both simulated and human data were split into training (50%), validation (25%), and testing (25%) sets. The human testing set did not include any data (even trials; $N = 1482$) used to estimate the generative fixation model parameters in Callaway et al. (2021). Of the trials not used to estimate model parameters (odd trials), half ($N = 742$) were randomly selected as the test set, which was held constant across runs. The remaining odd trials ($N = 742$) were combined with the even trials to form training and validation sets (2224 total). Within each run, 1482 trials (67%) were used for training, and 742 trials (33%) were for validation to select hyperparameters.

We trained Transformers (Vaswani et al., 2017), GRUs (Cho et al., 2014), and LSTMs (Hochreiter & Schmidhuber, 1997). Detailed descriptions of each model are provided in Supplementary section "Detailed information on model architectures." Because qualitative results were the same across architectures, we show only Transformer results in the main text and present results for all networks in the Supplementary Information (Figs. S2–S4). Control models used MLPs. All networks were implemented in the Python package, PyTorch (Paszke et al., 2017). We used the Adam optimizer to identify network parameters that minimized the mean square error in predicting the set of training sequences. All training used a batch size of 32. For each task, for all networks, we used a grid search to identify the number of hidden units (and embedding dimensionality for transformers; out of [8, 16, 32, 62, 128, 256, 512]) and learning rate (out of [.00001, .0001, .001]). For each combination of these hyperparameters, we trained five models, each with different starting weights. Resultant best hyperparameters are reported in Tables S1– S4.

When training on simulated data alone, networks were trained with a single epoch through 1.5 million simulated training trials. Models trained on human data alone and those pretrained on simulated data followed by fine-tuning on human data both used the same number of human training examples ($N = 1482$ trials). When training on human data alone, networks were trained for up to 1350 epochs through the training dataset of 1482 human trials. When pretraining on simulated data and then finetuning with human data, networks were trained with a single epoch through 1.5 million simulated training trials and then were finetuned for up to 1350 epochs through the training dataset of 1482 human trials. For all approaches, for each hyperparameter combination, we averaged the five error-versus-training-number curves, smoothed them with a Gaussian kernel ($\sigma = 200$ batches), and selected the hyperparameters and either the number of simulated training trials (for the simulated only case) or number of epochs through human data (for human only and simulated and human cases) that achieved minimum mean squared error under that approach and objective.

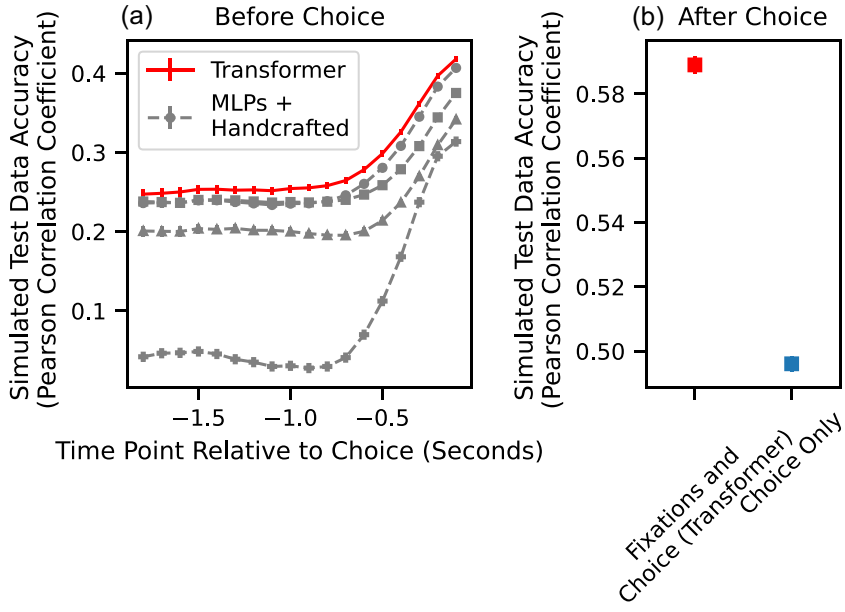

Fig. 4. Results of training model on simulated data and testing on held-out simulated data. (a) Predictive accuracy of neural networks at predicting simulated data utilities, at each time-point prior to a choice being made. Gray markers denote different control models that use MLPs to map handcrafted features to estimates of utility (Plus: current item, Triangle: proportion of time spent on each item, Square: cumulative time spent on each item, Circle: all prior features together). (b) Predictive accuracy after the choice is made. Transformers trained on simulated fixation and choice data outperform a model which only uses the choice that was made.

For the Transformer networks, we set the number of attention heads to 4 and the number of layers to 2. All other parameters were set to PyTorch default values. All final results reflect using these hyperparameters and number of training sequences, averaged over 100 runs, each with randomized training data ordering and initial weights.

4. Results

4.1. Transformers trained on simulated data can predict latent utilities

We first examined the ability of Transformers trained on simulated fixation and choice data to predict corresponding latent utilities used to generate that data. An advantage for predicting utilities from fixations in addition to choices, as opposed to predicting from choices alone, is that prediction from fixations can be made prior to the choice occurring. Indeed, Transformers were able to predict latent utilities at time-points prior to choice occurrence, from fixations alone, with prediction accuracy increasing up until the time of choice occurrence (Fig. 4a). This prediction accuracy prior to choice outperformed a variety of control models, which used MLPs to map hand-designed features at a single-time point to prediction of utilities (see Methods). The best-performing control model was provided the current fixation identity,

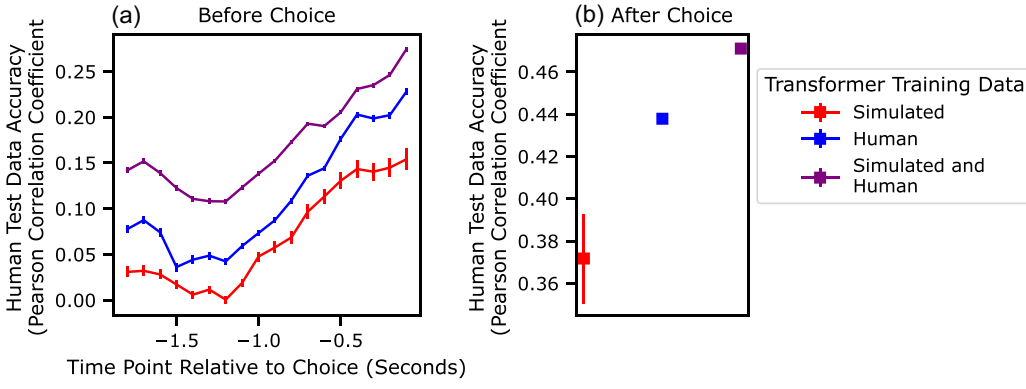

Fig. 5. Testing on human data under different training regimes. Transformers were trained using either simulated data alone, human data alone, or were pretrained with simulated data and finetuned with human data. Networks trained with both simulated and human data outperformed networks trained with either alone. (a) Predictive accuracy of neural networks at predicting self-reported human item utilities, at each time-point prior to a choice being made. (b) Predictive accuracy of neural networks at predicting utilities of human data after choice is made, using both fixation and choice information.

the sum of fixations on each item up to that time-point, and the proportion of fixations up to that time-point (Fig. S1A). This control model achieved worse prediction accuracy than the Transformers model (independent sample t-test comparing accuracy correlations aggregated across time-points, $t(358) = 19.2$, $p < .001$) demonstrating that the Transformers can learn nontrivial sequential aspects of the relationship between fixations and preferences in the simulated data (see Fig. S7 for additional demonstration that networks learn nontrivial sequential aspects of the relationship). Transformers trained on fixations in addition to choice also conferred an advantage in predicting preferences after a choice was made compared to predictions made using choice alone (Fig. 4B; $t(358) = 88.1$, $p < .001$). This demonstrates an ability to learn about relationships between fixations and preferences in simulated data beyond just predicting which item will be chosen.

4.2. Simulated data complements human data in predicting human utilities

We next sought to examine the ability of Transformers trained on fixation and choice data to predict human self-reported utilities of items from fixations and choices over those items. Additionally, we sought to determine whether training networks with simulated data provided a benefit over training with human data alone. We thus compared the ability of different Transformers to predict real human self-reported preferences, varying whether the Transformers were trained using simulated data only, human data only, or pretrained on simulated data and fine-tuned using human data. Networks pretrained on simulated data and finetuned with human data outperformed networks trained using either simulated data or human data alone, both when predicting preferences prior to a choice being made (Fig. 5a; Simulated and Human vs. Simulated Only: $t(358) = 25.6$, $p < .001$; Simulated and Human vs. Human Only: $t(358) = 30.8$, $p < .001$; see Fig. S6 for mean squared error of each

approach) and also when predicting with knowledge of the choice (Fig. 5b; Simulated and Human vs. Simulated Only: $t(358) = 9.2$, $p < .001$; Simulated and Human vs. Human Only: $t(358) = 45.4$, $p < .001$). This demonstrates that simulated data is beneficial in addition to human data in predicting real human preferences.

We additionally compared augmenting limited human data with simulated data to an alternative way limited human data might be augmented (Supplementary section "Comparison with noise-based data augmentation"). Specifically, we implemented a form of data augmentation by repeating existing human trials, but modifying each example with noise. We find that models trained on simulated-augmented data outperform those trained using noise-augmented data, further demonstrating the utility of pretraining with simulated data (Fig. S5).

To assess the overall accuracy at predicting human preferences from fixations alone, we compared Transformers trained on human and simulated data to control MLPs trained on hand-designed features. The best-performing control model for predicting human preferences from fixations was provided the proportion of fixations up to that time-point (Fig. S1B). We note that because the Transformer is trained to maximize accuracy across all time-points, it is outperformed at the final time-point prior to choice by the best control model (note, however, that neither model knows at what time-point the choice will occur). Aggregated across all time-points, however, this control model was outperformed by Transformers trained on simulated and human data (Fig. 6a; $t(358) = 37.3$, $p < .001$; see Fig. S8 for comparison against additional control model). As in the simulated data case, Transformers trained on simulated and human data, using both information about fixations and which item was chosen, performed better than a model that only used information about which item was chosen (Fig. 6b; $t(358) = 113.9$, $p < .001$). This demonstrates that, under this approach, using fixation data to predict preferences confers a slight benefit beyond simply predicting which item will be chosen.

5. Conclusion

Cognitive models, which define the relationships between an individual's latent preferences and their behavior, offer a tremendous opportunity to infer the hidden variables that guide an individual's choice. However, standard approaches to performing probabilistic inference with such models are computationally prohibitive for practical applications. Here, we have proposed and implemented a new approach for using neural networks to perform inference in such cognitive models, which can make inference computationally feasible for online applications. In addition to demonstrating that neural networks can perform inference of latent preferences in such models, we have also shown that simulating data from such models can make up for limited human data in training neural networks to infer real human preferences from behavior.

We found robust results across LSTMs, GRUs, and Transformers, but Transformers benefited most from pretraining with simulated data. This is consistent with recent work showing that Transformers especially benefit from massive pretraining datasets (e.g., Radford, Narasimhan, Salimans, & Sutskever, 2018). However, the similar qualitative effects

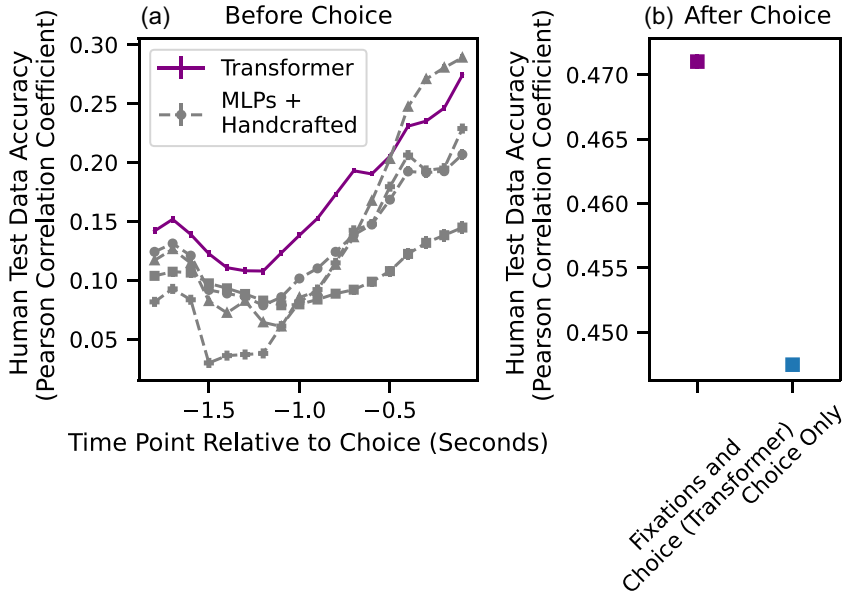

Fig. 6. Comparison of Transformers trained on simulated and human data to control model using hand-crafted features and also model which uses choice alone. (a) Predictive accuracy of neural networks at predicting self-reported human item utilities, at each time-point prior to a choice being made. Gray markers denote different control models that use MLPs to map handcrafted features to estimates of utility (Plus: current item, Triangle: proportion of time spent on each item, Square: cumulative time spent on each item, Circle: all prior features together). (b) Predictive accuracy on human data after the choice is made. Transformers trained on simulated and human fixation and choice data outperform a model which only uses the choice that was made.

across models suggest these are due to the data and any sufficiently capable architecture would show similar effects.

A key question for future analysis is what type of information neural network models utilize from the data-generative model. Callaway et al. (2021) demonstrate that the optimal fixation model captures a rich and subtle relationship between preference and gaze, which has been validated in human data. It predicts, for example, that the duration of the first fixation to an item, the order in which it is refixated, and whether it is fixated last should all be informative of preference. While we do not show the exact features that neural network models utilize to predictions, we do show in Supplementary analysis (Figs. S7 and S8) that preserving the order of fixations and time-points is important for model performance, demonstrating that it picks up on fine-grained relationships between fixations and choice.

Overall, this approach is likely limited by the extent to which cognitive models can capture idiosyncratic features of the relationship between human preferences and behavior. In future work, we can improve this approach by identifying and understanding discrepancies between model generated datasets and real human fixation data. Identifying such discrepancies may enable the generation of new generative models of fixations. These models may relax the

strong optimality assumptions of the model we currently use, but may in turn produce fixation data that is more useful for training neural networks for predicting preferences.

Code availability statement

Analysis code is available at https://github.com/evanrussek/Inverse_Gaze_Neural_Networks.

Acknowledgments

This work was supported by Facebook Reality Labs (now Meta).

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(16), 8825–8835.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bourgin, D. D., Peterson, J. C., Reichman, D., Griffiths, T. L., & Russell, S. J. (2019). Cognitive model priors for predicting human decisions. In *Proceedings of the 36th International Conference on Machine Learning*, volume *97* (pp. 5133–5141).

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology*, *17*(3), 1–29.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111). Association for Computational Linguistics.

Fengler, A., Govindarajan, L. N., Chen, T., & Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *eLife*, *10*: e65074.

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, *217*, 104885.

Ger, Y., Nachmani, E., Wolf, L., & Shahar, N. (2023). Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior. *bioRxiv*, 2023.04.21.537666.

Glaholt, M. G., Wu, M.-C., & Reingold, E. M. (2009). Predicting preference from fixations. *PsychNology Journal*, *7*(2), 141–158.

Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634–645.

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., Greenberg, D. S., & Macke, J. H. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLlife*, *9*: e56261.

Goyal, S., Miyapuram, K. P., & Lahiri, U. (2015). Predicting consumer's behavior using eye tracking data. In *2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI)* (pp. 126–129).

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29* (pp. 3909–3917).

Ho, M. K., & Griffiths, T. L. (2022). Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, *5*(1), 33–53.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jang, A. I., Sharma, R., & Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *eLlife*, *10*: e63436.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108* (33), 13852–13857.

Kuperwajs, I., Schütt, H. H., & Ma, W. J. (2023). Using deep neural networks as a guide for modeling human planning.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS One*, *9*(3), e92160.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, ICML '00 (pp. 663–670). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Papamakarios, G., & Murray, I. (2016). Fast-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems 29* (pp. 1028–1036).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017).Automatic differentiation in PyTorch. In *NPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Rabinowitz, N. C., Perbet, F., Francis Song, H., Zhang, C., Ali Eslami, S. M., & Botvinick, M. (2018). Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume *80* (pp. 4215–4224).

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Kothe, U. (2022). BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(4), 1452–1466.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *Open AI Blog*.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Sequential thought processes in PDP models. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, *2*, 3–57.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, volume *30* (pp. 5998–6008).

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10). Sci. Adv.6, eaax5979

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Performance of nonsequential control models.

Fig. S2. Results of training model on simulated data and testing on held-out simulated data for GRU and LSTM networks.

Fig. S3. Testing on human data under different training regimes for GRU and LSTM networks.

Fig. S4. Results of training model on simulated and human data and testing on held-out human data for GRU and Transformer networks.

Fig. S5. Comparison of models trained with alternative data augmentation approaches to models pretrained on simulated data.

Fig. S6. Mean squared error for models trained on human data only and models pretrained on simulated data.

Fig. S7. Comparison of models trained on shuffled fixations up to some time-point with those trained on non-shuffled fixations.

Fig. S8. Comparison of models trained on shuffled fixations up-to some time-point with those trained on non-shuffled fixations.

Table S1. Hyperparameters for training on simulated data and predicting simulated data.

Table S2. Hyperparameters for training on simulated data and predicting human data.

Table S3. Hyperparameters for training on human data and testing on human data.

Table S4. Hyperparameters for training on simulated and human data and testing on human data.