Cognitive Model Priors for Predicting Human Decisions

David D. Bourgin [*1] **Joshua C. Peterson** [*2] **Daniel Reichman** [2] **Stuart J. Russell** [1] **Thomas L. Griffiths** [2]

Abstract

Human decision-making underlies all economic behavior. For the past four decades, human decision-making under uncertainty has continued to be explained by theoretical models based on prospect theory, a framework that was awarded the Nobel Prize in Economic Sciences. However, theoretical models of this kind have developed slowly, and robust, high-precision predictive models of human decisions remain a challenge. While machine learning is a natural candidate for solving these problems, it is currently unclear to what extent it can improve predictions obtained by current theories. We argue that this is mainly due to data scarcity, since noisy human behavior requires massive sample sizes to be accurately captured by off-the-shelf machine learning methods. To solve this problem, what is needed are machine learning models with appropriate inductive biases for capturing human behavior, and larger datasets. We offer two contributions towards this end: first, we construct "cognitive model priors" by pretraining neural networks with synthetic data generated by cognitive models (i.e., theoretical models developed by cognitive psychologists). We find that fine-tuning these networks on small datasets of real human decisions results in unprecedented state-of-the-art improvements on two benchmark datasets. Second, we present the first large-scale dataset for human decision-making, containing over 240,000 human judgments across over 13,000 decision problems. This dataset reveals the circumstances where cognitive model priors are useful, and provides a new standard for benchmarking prediction of human decisions under uncertainty.

[*]Equal contribution [1]University of California, Berkeley [2]Princeton University. Correspondence to: David D. Bourgin <ddbourgin@gmail.com>, Joshua C. Peterson <peterson.c.joshua@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

1. Introduction

Which gamble would you rather take: a 50/50 chance of winning/losing $100, or a 100% chance of $0? Although both gambles have equal expected value (payoff), the majority of people systematically prefer the second (Kahneman & Tversky, 1979). Predicting the choices that people make in such situations is of central importance in economics and relevant to understanding consumer behavior. As suggested above, a recurring empirical finding is that people reliably deviate from optimal decision-making under various conditions. Extensive research has sought to understand these deviations and the conditions that elicit them (Edwards 1954; Baron 2000; Gilovich et al. 2002, to name just a few). A prime example is prospect theory, devised by Kahneman & Tversky (1979), and a key part of Kahneman's receipt of the Nobel Prize in Economic Sciences. Constructing models of choice behavior based on prospect theory and its descendants (e.g., Tversky & Kahneman 1992; Erev et al. 2017) remains an active research area in behavioral economics and cognitive psychology. Good predictive models of human choice also arise in AI applications as they can be used to build artificial agents that are better aligned with human preferences (Rosenfeld & Kraus, 2018; Russell & Norvig, 2016; Chajewska et al., 2001).

Despite these efforts, we are still far from accurate and robust models of how people make decisions. Theoretical explanations for deviations from maximizing expected value are often contradictory, making it difficult to come up with a single framework that explains the plethora of empirically observed deviations such as loss aversion, the St. Petersburg paradox, and many others (Erev et al., 2017). This state of affairs has motivated several recent choice prediction competitions (Erev et al., 2010; 2017; Plonsky et al., 2019) in an attempt to achieve better predictive models of human decisions.

One approach towards improving prediction has been to leverage machine learning (ML) (Rosenfeld & Kraus, 2018; Plonsky et al., 2017; Noti et al., 2017; Plonsky et al., 2019). However, data scarcity has remained the key limiting factor in leveraging ML to predict human decisions. Cognitive models of human decision-making (e.g., prospect theory) are typically developed by behavioral scientists using datasets consisting of hundreds or fewer data points. Such

small datasets, coupled with the complex and noisy nature of human behavior, increase the risk of overfitting and limit the applicability of machine learning methods (Geman et al., 1992; Yarkoni & Westfall, 2017).

In this work we develop and evaluate a method for enabling machine learning to achieve better predictions of human choice behavior in settings where data are scarce. Our approach is to treat cognitive models as a source of inductive bias to help machine learning methods "get off the ground," since their predictions are closer to human behavior than untrained machine learning models. The resulting "cognitive model priors" offer a solution to the problem of prediction under data scarcity by combining the rich knowledge available in cognitive models with the flexibility to easily adapt to new data.

Contributions: Our contributions are as follows:

- We introduce a simple and general methodology for translating cognitive models into inductive biases for machine learning methods. At a high level, our method consists of generating *synthetic datasets* generated from cognitive models that can be used to establish informative priors for ML algorithms. We focus on the case where a neural network is trained on these synthetic datasets and then fine-tuned using a much smaller dataset of real human data.
- We test this methodology using two new large synthetic datasets (`synth15` and `synth18`) that capture current theoretical knowledge about human decision-making. Transferred to ML models, the resulting cognitive model priors provide a new way to predict human decisions that integrates the flexibility of ML with existing psychological theory.
- Using this new methodology we greatly improve state-of-the-art performance on two recent human choice prediction competition datasets (Erev et al., 2017; Plonsky et al., 2019). The resulting models operate on raw features alone (not relying on hand-designed, theory-based features), for the first time for a competitive machine learning model of human decisions.
- We introduce a new benchmark dataset for human choice prediction in machine learning that is an order of magnitude larger than any previous datasets, comprising more than 240,000 human judgments over 13,000 unique decision problems. We show that even in datasets of this magnitude, cognitive model priors reduce prediction error and increase training efficiency.

Related Work: There has been a growing interest in applying machine learning to problems in behavioral economics (Camerer, 2018; Peysakhovich & Naecker, 2017; Hartford et al., 2016) and econometrics (Mullainathan & Spiess, 2017). Several recent works have leveraged ML to gain new theoretical insights (Peysakhovich & Naecker, 2017; Fudenberg & Liang, Accepted; Kleinberg et al., 2017), while other work has focused solely on improving prediction. For example, Hartford et al. (2016) introduced an inductive bias in the form of a specialized neural network layer, allowing for improved prediction in the face of scarce training data. The particular bias proposed, however, was only applicable to the domain of two-player games.

Most closely related to our work is that of Plonsky et al. (2017), who applied an array of popular ML algorithms to the problem of predicting human decisions for pairs of gambles (the same gambles we will consider in this paper). In addition to providing baseline performance results for a number of ML algorithms on human decision data, the authors used significantly more data than was present in previous work (e.g., Peysakhovich & Naecker, 2017). Notably, the authors found that the predictions made using these algorithms were much poorer than a baseline model based on psychological theory when the raw gambles were used as input. Interestingly, however, when model inputs were supplemented by the components of the theoretical model (expressed as features), one of the ML models—a random forest algorithm—showed an improvement over the baseline psychological model. The authors also found that using theory-based predictions as additional inputs improves performance even further (Plonsky et al., 2017; 2019). These approaches differ from ours in that they require a model-to-feature decomposition in order to leverage the theoretical model, which may not be unique, and can require significant effort and expert knowledge.

2. Decision-Making Under Uncertainty

Human decision-making under uncertainty has traditionally been studied in the context of sets of gambles with uncertain outcomes. A single gamble is a set of N possible outcomes x_i and their probabilities p_i. Given a set of M gambles, the decision-maker must choose the single most appealing option. In the remainder of the paper, we will be concerned with learning a model that can predict the probability that a human decision-maker will choose each gamble. We want to infer a probability as opposed to a binary selection in order to capture variability both within a single person's choice behavior and across different people.

2.1. Cognitive Models of Decision-Making

Given a utility function $u(\cdot)$ that reflects the value of each outcome to a decision-maker, the rational solution to the choice problem is to simply choose the gamble that maximizes expected utility (EU): $\sum_{i=1}^{N} p_i u(x_i)$ (Von Neumann & Morgenstern, 1944). Interestingly, humans do not appear to adhere to this strategy, even when allowing for a

range of utility functions. In fact, cognitive models of human decision-making were originally designed in order to capture four *deviations* from EU, with the most influential model being prospect theory (Kahneman & Tversky, 1979). This model asserts that people assess a quantity that takes a similar form to expected utility, $V = \sum_{i=1}^{N} \pi(p_i)v(x_i)$, but where $\pi(\cdot)$ is a weighting function that nonlinearly transforms outcome probabilities and $v(\cdot)$ is a subjective assessment that can be influenced by factors such as whether the outcome is perceived as a gain or a loss.

Since prospect theory was first introduced, the list of discovered human deviations from expected utility has grown significantly, making it increasingly difficult to capture them all within a single unifying model (Erev et al., 2017). One of the most successful recent attempts to do so is the Best Estimate and Sampling Tools (BEAST) model, which eschews subjective weighing functions in favor of a complex process of mental sampling. BEAST represents the overall value of each prospect as the sum of the best estimate of its expected value and that of a set of sampling tools that correspond to four behavioral tendencies. As a result, gamble A will be strictly preferred to gamble B if and only if:

$$[BEV_{\mathrm{A}} - BEV_{\mathrm{B}}] + [ST_{\mathrm{A}} - ST_{\mathrm{B}}] + e > 0, \quad (1)$$

where BEV_{A}–BEV_{B} is the advantage of gamble A over gamble B based on their expected values, ST_{A}–ST_{B} is the advantage based on alternative sampling tools, and e is a normal error term.

Sampling tools include both biased and unbiased sample-based estimators designed to explain four psychological biases: (1) the tendency to assume the worst outcome of each gamble, (2) the tendency to weight all outcomes as equally likely, (3) sensitivity to the sign of the reward, and (4) the tendency to select the gamble that minimizes the probability of immediate regret.[1] In the remainder of the paper, we use BEAST as a proxy for current theoretical progress as it both explains a large number of behavioral phenomenon and was built explicitly to address concerns about the applicability of theoretical models to robust prediction.

2.2. Choice Prediction Competitions

Two important resources for evaluating predictive models of human decision-making are the 2015 and 2018 Choice Prediction Competition datasets (CPC15 and CPC18, respectively; Erev et al., 2017; Plonsky et al., 2019). These competitions offer two benefits. First, they encompass a large space of sets of gambles. Whereas behavioral experiments tend to study only one phenomenon at a time, the CPC datasets were explicitly designed to include sets of gambles that elicit all known deviations from EU in addition to other gambles sampled from a much larger problem space. Second, although we will later argue that such datasets are still too small to fully train and evaluate predictive models, they are currently the largest of their kind. CPC15 contains 90 choice problems (30 test problems) for five repeated trials for a total of 450 datapoints (150 test datapoints), and CPC18 (a superset of CPC15) contains 210 choice problems (90 test problems) for five repeated trials for a total of 1,050 datapoints (450 test datapoints).

Problems in the CPC datasets required people to choose between a pair of gambles, A and B. Each gamble consisted of a collection of rewards and their associated outcome probabilities. In CPC15, gamble A was constrained to only have only two outcomes (similar to the example given in section 2). Gamble B yielded a fixed reward with probability $1 - p_L$ and the outcome of a lottery (i.e., the outcome of another explicitly described gamble) otherwise. That is, by convention in the competition problems, a lottery is defined as the outcome of a chosen gamble (occurring with probability p_L) that can also yield one of multiple monetary outcomes, each with some probability. Gamble B's lottery varied by problem and was parameterized using a range of options, including the number of outcomes and the shape of the payoff distribution. In CPC18, the only difference was that some of the problems allowed gamble A to take on the more complex lottery structure of gamble B.

For each problem, human participants made sequential binary choices between gamble A and B for five blocks of five trials each. The first block of each problem was assigned to a no-feedback condition, where subjects were shown only the reward they received from the gamble they selected. In contrast, during the remaining four blocks subjects were shown the reward they obtained from their selection as well as the reward they could have obtained had they selected the alternative gamble. Finally, gambles for problems assigned to an "ambiguous" condition had their outcome probabilities hidden from participants. Further details on the design and format of the gamble reward distributions can be found in Erev et al. (2017) and Plonsky et al. (2019). Prediction of the aggregated human selection frequencies (proportions between 0 and 1) is made given a 12-dimensional vector of the parameters of the two gambles and the block number. As this information is all displayed to the participant in the course of their selection, we refer to these as "raw" problem features. Prediction accuracy is measured using mean squared error (MSE).

To summarize, both CPC15 and CPC18 contain choice problems separated into training and test sets. The task is to build or train a model to predict the probability (i.e., the proportion of human participants) that selected gamble A for each problem based on the training set, and evaluate on the test set using MSE.

[1] For further details, see Erev et al. (2017), and source code at: `cpc-18.com/baseline-models-and-source-code/`

2.3. Data Scarcity in Behavioral Sciences

Scientific studies with human participants often yield small datasets. One reason for this is resource limitations: compensating participants is costly, and large-scale data collection can be limited by the ability to recruit participants. A second reason is that small datasets are a byproduct of the theory-building process: because it is rare to know all the relevant variables ahead of time, researchers must isolate, manipulate, and analyze manageable sets of independent variables over many studies rather than collect a single large dataset.

The high variability of human behavior also makes prediction challenging. Indeed, many important types of human behavior (e.g., decision making) differ significantly across individuals and require large sample sizes to accurately assess. For example, the number of problems in CPC15 and CPC18 for which human data could be feasibly collected was limited because obtaining stable estimates of aggregate behavior required a large number of human participants per choice problem.

Small datasets and high variability may help to explain why machine learning approaches have garnered only marginal improvements in predicting human decisions. Given these challenges, it is worth considering possible sources of human-relevant inductive biases that might help alleviate the symptoms of data scarcity.

3. Cognitive Model Priors

Scientific theory-building is principally concerned with *understanding* and *explanation*, whereas machine learning is geared more towards *prediction*, often employing complex and opaque models. Given enough data, and assuming the domain is of tractable complexity, a machine learning model might rediscover some of the components of an established scientific theory, along with a number of new, potentially complicated and nuanced improvements that may be hard to immediately explain, but are responsible for robust and accurate forecasting.

When data is scarce, however, theoretical models (which tend to be relatively parsimonious compared to models from ML) offer a valuable source of bias, resulting in inferior accuracy but potentially superior generalization (Yarkoni & Westfall, 2017). Ideally, we would like to find a way to leverage the hard-won generalizability of psychological theories while making use of the flexibility of machine learning methods to develop more powerful predictive models.

Towards this goal, we propose building a bridge between the two modeling domains via the common currency of data. Assuming a cognitive model can be expressed as some function $f(\cdot)$, where inputs are experimental task descriptions or stimuli and outputs are human choices, preferences, or behaviors, we can convert the insights contained in the scientific model into a synthetic dataset and new model as follows:

1. Evaluate a large range of inputs, including those without accompanying human targets, to generate input-target pairs (x_i, $f(x_i)$) for training.

2. Train a machine learning model to approximate $f(\cdot)$ via (x_i,$f(x_i)$) as opposed to approximating human decision functions directly (x_i and human targets h_i).

3. Fine-tune the resulting model on small, real human datasets (x_i,h_i) in order to learn fine-grained improvements to the scientific model.

By instantiating theoretical knowledge in a synthetic dataset, we provide a domain-general way to construct a *cognitive model prior* for ML algorithms. This prior can be explicit (conjugate priors for Bayesian models can be interpreted as encoding the sufficient statistics of previous experience (Box & Tiao, 2011)) or implicit (if used for pretraining neural networks, those networks will be regularized towards pretrained weights rather than an arbitrary set of initial weights).

3.1. Converting Choice Prediction Models to Data

We sampled approximately 100k new problems from the space of possible CPC15 problems and 85k problems from the space of possible CPC18 problems using the procedures from Plonsky et al. (2017) (Appendix D) and Plonsky et al. (2018) (Appendix D). These datasets of sampled problems are at least two orders of magnitude larger than those generated previously using these procedures meant for human experiments. We found these dataset sizes to be both feasible to generate and sufficient to allow off-the-shelf ML methods to learn good approximations to cognitive models. For each new dataset we ensured that no problem occurred more than once or overlapped with the existing CPC15 or CPC18 problems, and removed "degenerate" problems in which either both gambles had the same distribution and payoffs or at least one gamble had no variance, but the rewards for the two gambles were correlated. We denote these new collections of problems as the `synth15` and `synth18` datasets, respectively.

To create training targets for our cognitive model prior, we used two publicly available versions of the BEAST model (section 2.1. These include the original BEAST model proposed before CPC15, which we denote as `BEAST15`, and the subsequent "Subjective Dominance" variant introduced before CPC18, which we denote as `BEAST18`. These two models were used to predict gamble selection frequencies for each problem in the `synth15` and `synth18` datasets

respectively. These predictions could then be used as training targets for our machine learning model in order to transfer insights in the theoretical models into a form that can be easily modified given new data to learn from.

3.2. Model Setup and Fine-Tuning

We opted to use neural networks to transfer knowledge from BEAST because they can be easily fine-tuned to real human data after initial training and provide a number of different options for regularization. While sequential learning of this sort risks catastrophic forgetting (French, 1999), we found that fine-tuning with a small learning rate (1e-6) was sufficient to make use of prior knowledge about BEAST internalized in the network's weights. This allows for a straightforward integration of cognitive model priors into the standard neural network training paradigm.

Neural Network Architecture. We grid-searched approximately 20,000 hyperparameter settings and found the best multilayer perceptron (MLP) overall for estimating both variants of BEAST (as well as the other datasets/tasks in this paper) had three layers with 200, 275, and 100 units respectively, SReLU activation functions, layer-wise dropout rates of 0.15, and an RMSProp optimizer with a 0.001 learning rate. The output layer was one-dimensional with a sigmoid activation function to match the range of the human targets. We also obtained lower error, less overfitting, and more stable fine-tuning using the SET algorithm (Mocanu et al., 2018), which consists of (1) initializing the network as an Erdős–Rényi random sparse graph and (2) evolving this graph at the end of each epoch via both pruning of small weights and the addition of new random connections. While we don't view this architectural choice as essential for the application of cognitive model priors, this form of model compression may be additionally useful when data is scarce.

3.3. Results

CPC15: The results of our analysis along with the performance of several baseline and competing models (drawing on Plonsky et al. 2017) are given in Table 1. For example, a wide range of common machine learning algorithms utterly fail to obtain reasonable MSE scores given the raw gamble parameters shown to human participants, likely due to a lack of data. For CPC15, `BEAST15` was the competition's theoretical baseline model, and beats all unaided ML methods. The competition organizers also note that no other theory-based models from psychology or behavioral economics, such as prospect theory models, reached the leaderboard's top-10.[2] The CPC15 winner was a small augmentation of `BEAST15` with no machine learning component. The table section *ML + Feature Engineering* shows results for the method employed in Plonsky et al. (2017), wherein intermediate features derived from `BEAST15` were used as input to ML algorithms as opposed to raw features alone (i.e., parameters of the gambles). Notably, these models also do worse than the self-contained `BEAST15` model with the exception of the random forest model, which lowered test set MSE of the CPC15 winner by 0.0001. More successful was an ensemble that included the random forest model with the `BEAST15` prediction as an additional feature, obtaining an MSE of 0.007. In contrast to all of these baselines, our own method of fine-tuning a neural translation of `BEAST15` by way of `synth15` obtains a large reduction in MSE to 0.0053 (24% decrease).

Table 1. Performance (MSE) for CPC15 and CPC18 benchmarks.

	Model	MSE×100
CPC 2015	*ML + Raw Data*	
	MLP	7.39
	k-Nearest Neighbors	7.15
	Kernel SVM	5.52
	Random Forest	6.13
	Theoretical Models	
	`BEAST15`	0.99
	CPC 2015 Winner	0.88
	ML + Feature Engineering	
	MLP	1.81
	k-Nearest Neighbors	1.62
	Kernel SVM	1.01
	Random Forest	0.87
	Ensemble	0.70
	MLP + Cognitive Prior (ours)	**0.53**
CPC 2018	*Theoretical Models*	
	`BEAST18`	0.70
	ML + Feature Engineering	
	Random Forest	0.68
	CPC 2018 Winner	0.57
	MLP + Cognitive Prior (ours)	**0.48**

CPC18: While the same set of baselines are not available for CPC18, the most crucial comparisons are given in the second half of Table 1. First, `BEAST18`, a modification of `BEAST15`, obtains a fairly low MSE score of 0.007. Since CPC15 is a subset of CPC18, this can be considered an improvement. Second, Plonsky et al. (2018) released a random forest baseline that makes use of theory-inspired input features, given it was the only successful application of ML in CPC15. Like CPC15, a small decrease in MSE of 0.0002 is obtained. The winner of the CPC18 competition—a submission by the authors of the current paper—was a gradient-boosted decision tree regressor and obtained a notable decrease in MSE for the first time using ML methods. Part of this success is likely due to the larger CPC18 train-

[2] `http://departments.agri.huji.ac.il/economics/teachers/ert_eyal/compres.htm`

ing set. Finally, the further improvement we provide in this paper is a `synth18`-trained MLP that was fine-tuned on the CPC18 training set that produced an MSE of 0.0048, a decrease in error of 0.0022 (16%) over the competition's theoretical baseline. The resulting model (as well as our MLP for CPC15) take only raw gamble parameters as input when fully trained.

4. A New Human Decision Dataset

In many ML competitions, the published test sets often begin to function as unofficial validation sets, eventually leading to validation overfitting (Recht et al., 2018). Further, CPC test sets are much smaller than most benchmark datasets in machine learning, making for a poor overall test of generalization. Thus, even though cognitive model priors may make it possible to train on small datasets, a larger dataset is still necessary to establish their value with confidence. In the following section we introduce `choices13k`, a new large-scale dataset of human risky-choice behavior, and use it to evaluate our cognitive model priors. Before doing so, however, we preview a crucial initial result on this dataset to underscore the risk of using small test sets. In Figure 1 we drew one-hundred bootstrap samples from our larger dataset to simulate variability in small samples of validation problems. When drawing samples the size of CPC18, the variability in fit by `BEAST18` is large, and stabilizes considerably when taking samples of up to approximately 6,500 problems. This motivates the need to obtain a larger validation set for assessing the utility of cognitive model priors, and also provides an opportunity to assess whether such priors have a benefit even when data is in abundance.

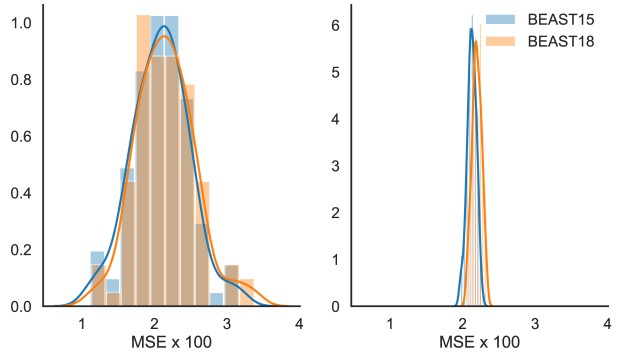

Figure 1. **Left:** The distribution of BEAST MSEs over 100 bootstrapped samples of size equal to the CPC 2018 training dataset (210 problems). **Right:** The distribution of BEAST MSEs on 50% of the entire `choices13k` dataset (approx. 6,500 problems).

4.1. Data Collection for `choices13k`

We collected a dataset more than ten times the size of CPC18 using Amazon Mechanical Turk. In total, the new dataset

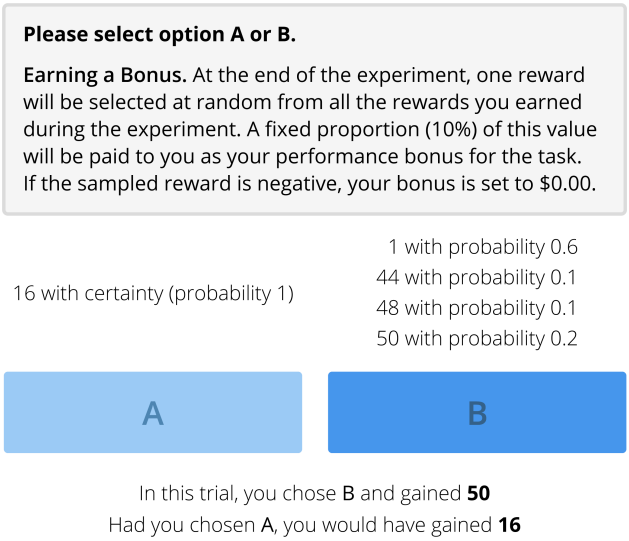

Figure 2. Experiment interface for collecting human gamble selections, modeled after the interface in Erev et al. (2017).

contained 242,879 human judgments on 13,006 gamble selection problems, making it the largest public dataset of human risky choice behavior to date. The gamble selection problems were sampled uniformly at random from the `synth15` dataset. The presentation format for the gambling problems was inspired by the framework used in the 2015 and 2018 Choice Prediction Competitions (Erev et al., 2017; Plonsky et al., 2019). The only exceptions in our datasets is that the `block` parameter in CPC15 and CPC18 was reduced from five (ie., subjects completed five blocks of five trials for each problem) to two, and that feedback and no-feedback blocks were not required to be presented sequentially. We found that this alteration did not significantly affect overall predictive accuracy in our models, and allowed us to more than halve the number of trials necessary for each problem.

As in the CPC datasets, each problem required participants to choose between two gambles, after which a reward was sampled from the selected gamble's payoff distribution and added to the participant's cumulative reward (Figure 2). Participants that selected gambles on the same side of the screen over more than 80% of the trials were excluded. In all, the final dataset consisted of gamble selection frequencies for an average of 16 participants per problem. Each participant completed 16 problems in the feedback condition and four problems in the no feedback condition and was paid $0.75 plus a bonus of 10% of their winnings from a randomly selected problem.

4.2. Assessing the Value of Cognitive Model Priors

The expanded `choices13k` dataset allows us to quantify the predictive power of psychological theory, and to assess

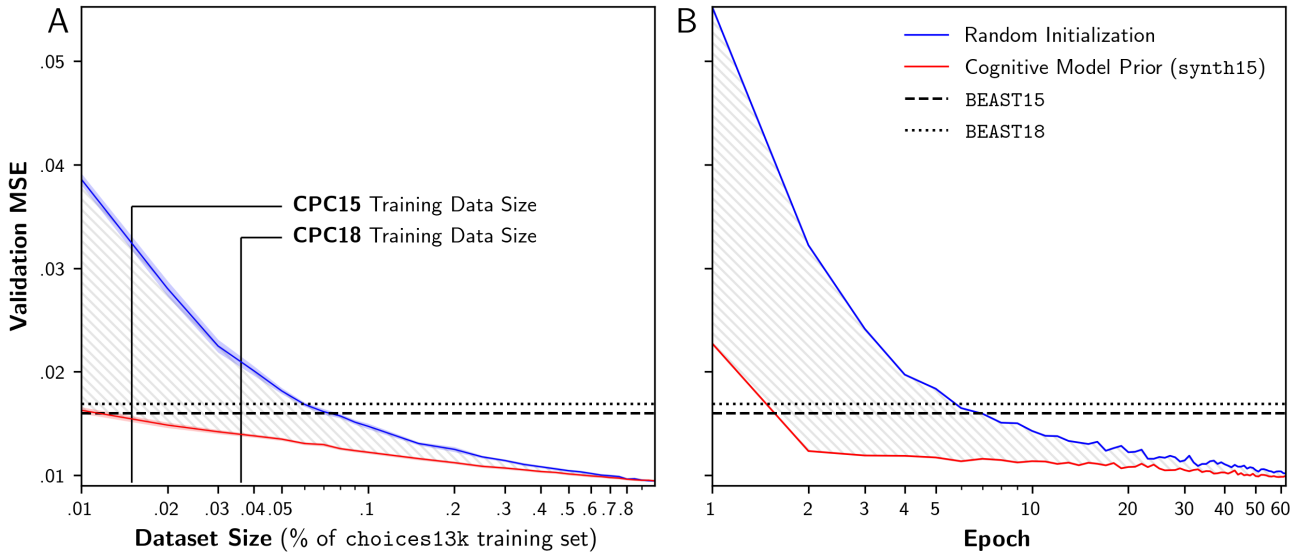

Figure 3. **A.** Validation MSE (20% of full `choices13k` human dataset) as a function of training on proportions of the training set (80% of full `choices13k` human dataset), for sparse MLPs trained from a random initialization (blue), and with a cognitive model prior (red). The prior allows for comparable MSE with significantly less data. **B.** Validation MSE (20% of full `choices13k` human dataset) as a function of training epoch. Using a prior allows for faster training. Both x-axes are shown in log-scale.

the relationship between data scarcity and the influence of our cognitive model prior. To this end, we varied the amount of training data available to neural networks that were either randomly initialized or pretrained on the `synth15` cognitive model prior. The training sets we varied were a proportion (from 0.01 to 1.0) of our full `choices13k` training set (80% of the overall dataset). The remaining 20% of `choices13k` was used as a constant validation set and the size did not vary. We repeated this process ten times.

Figure 3A plots validation MSE for a sparse MLP trained from a random initialization (blue) and one pretrained on `synth15` (i.e., using a cognitive model prior). Standard error contours of the average MSE scores, computed over ten train/validation splits, are shown in lighter colors. Note that models were re-initialized for each split and so variation due to initialization is a subset of the reported error contour regions. The MLP with a cognitive model prior begins with a significantly lower MSE in comparison to the randomly initialized model (0.016 versus 0.039), and continues to exhibit better error scores all the way up to 100% of the `choices13k` training set (0.00945 versus 0.00947). The advantage at 25% of the training data is a decrease by 0.0009, the advantage at 50% is a decrease by 0.0003, and the advantage at 75% is a decrease by 0.0002. It takes just 4% of our real human data (an amount still larger than CPC18) to cross over from an MSE above the best obtained by all other types of machine learning algorithms we tried (see section 4.3), to one below it, a reduction from 0.020 to 0.0138. These results suggests that cognitive model priors may be useful in any setting where collecting human decisions is costly, and further supports the pattern of results in section 3.3.

Beyond overall error, Figure 3B shows validation MSE per epoch given the full `choices13k` training dataset. Even though MSE for the two types of models largely converges (with a lingering numerical advantage in final MSE when using the prior), the MSE for the MLP with a cognitive model prior starts lower, decreases much faster, and requires fewer overall epochs to converge. This suggests one advantage of cognitive model priors that we did not anticipate: faster training. While datasets of this size do not require models that take as long to train as, e.g., large deep neural networks, we found this property extremely useful in allowing for larger hyperparameter grid searches.

4.3. A Benchmark for Predicting Human Decisions

Given the size of our dataset relative to previous ones, we propose it as the new benchmark for validating the precision and generalization of both theoretical and machine learning models. To this end, we provide a set of starting baseline scores as a challenge to ML practitioners (see Table 2). Notably, while the authors of Plonsky et al. (2017) achieved the best results on their dataset using random forest models, we found neural networks were much more effective for our larger dataset. Our most successful model, the result of a grid-search over 20,000 hyperparameter combinations, was a sparse MLP as described above, containing less than 10,000 parameters (down from around 100,000 for the best full dense MLP). By the end of training (using all

choices13k data), the synth15 cognitive model prior only added a slight numerical advantage in MSE. The advantage gained by enforcing sparsity likely indicates that our dataset, while much larger than any other of its kind, may still benefit from additional methods for accommodating data scarcity. Given the relatively generic (i.e., off-the-shelf) models used here, we expect there is much room for improvement in MSE on our dataset, particularly with specialized neural architectures combined with new ways of exploiting our cognitive model priors.

Table 2. Baseline validation MSE scores on the choices13k benchmark, averaged over ten 80/20 splits.

Model	MSE×100
Linear Regression	4.02
k-Nearest Neighbors	2.27
Kernel SVM	2.16
Random Forest	1.41
MLP	1.03
Sparse MLP (Mocanu et al., 2018)	**0.91**

5. Discussion

In this paper we have provided a method for addressing the data scarcity problem that limits the application of machine learning methods to small behavioral datasets by using cognitive models as a source of informative priors. We have outlined a novel approach for doing just that: train a neural network on a large synthetic dataset generated from a cognitive model and use the resulting weights of the model as a *cognitive model prior* that is then fine-tuned on additional human data. Following this approach we were able to efficiently outperform previously proposed cognitive and machine learning methods in predicting human decisions. We believe that the ideas suggested here may prove useful to other behavioral domains such as social cognition, marketing, and cognitive neuroscience.

The weights of the pretrained networks that encode the cognitive model priors could potentially be useful even in broader contexts if they encode a general value function for gambles. For example, one could imagine adding a new input layer with $M > 2$ gambles to choose from, which could be trained in isolation before fine-tuning the full network.

The utility of cognitive model priors for human choice prediction raises a number of questions. For example, how "complex" must the cognitive model prior be in order to produce a marked improvement in prediction error? Does taking EU or prospect theory as the source of a prior suffice to improve upon state-of-the-art prediction? Beyond theoretical interest, this question might have practical applications as it may indicate how much effort to invest in theory-driven models before they could be used as a source of priors. Further, our application of the prior was in the form of weight initialization (after pretraining), but there are many other possibilities, such as an auxiliary loss term for BEAST targets during fine-tuning, or an L2 weight prior centered on the initial weights. Both of these modifications may improve generalization.

Finally, although our focus here has been on prediction, there are many other important ways in which models of human decision-making are used (e.g., providing explanations, interpreting behavior, inferring causality, etc). We do not claim that models of human behavior should be evaluated based solely on their predictive accuracy. Indeed, the current approach trades some of the interpretability of the theoretical models in Plonsky et al. (2019) for the ease of integrating that theoretical knowledge into a more powerful predictive modeling framework. However, we do believe that better predictive models, even those driven by data rather than theory, can improve our understanding of human behavior: if machine learning methods can reach or exceed the predictive power of psychological theories, they may be able to point theoreticians toward regularities in behavior that have not yet been detected by human scientists.

6. Conclusion

In the current paper we introduced an approach to incorporating theory-based inductive biases from a cognitive model into machine learning algorithms. We demonstrated this approach with two new synthetic datasets: synth15 and synth18, and showed that that when data is in short supply, the resulting cognitive model priors allowed our network to achieve significantly better generalization performance with fewer training iterations than equivalent models with no such priors. We also found that integrating these cognitive model priors into a generic neural network model improved state-of-the-art predictive performance on two public benchmarks, CPC15 and CPC18, without relying on hand-tuned features. Finally, we introduced choices13k, the largest public dataset of human risky choice behavior to date. It is our hope that these three new datasets will support further interaction between the machine learning community and the behavioral sciences, leading to better predictive models of human decision-making.

Acknowledgments

This work was supported by grants from the Future of Life Institute, the Open Philanthropy Foundation, DARPA (cooperative agreement D17AC00004), and grant number 1718550 from the National Science Foundation.

References

Baron, J. *Thinking and deciding*. Cambridge University Press, 2000.

Box, G. E. and Tiao, G. C. *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011.

Camerer, C. F. Artificial intelligence and behavioral economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018.

Chajewska, U., Koller, D., and Ormoneit, D. Learning an agent's utility function by observing behavior. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 35–42. Morgan Kaufmann Publishers Inc., 2001.

Edwards, W. The theory of decision making. *Psychological Bulletin*, 51(4):380, 1954.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., Stewart, T., West, R., and Lebiere, C. A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23(1):15–47, 2010.

Erev, I., Ert, E., Plonsky, O., Cohen, D., and Cohen, O. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, 124(4):369–409, 2017.

French, R. M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Fudenberg, D. and Liang, A. Predicting and understanding initial play. *American Economic Review*, Accepted.

Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Gilovich, T., Griffin, D., and Kahneman, D. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

Hartford, J. S., Wright, J. R., and Leyton-Brown, K. Deep learning for predicting human strategic behavior. In *Advances in Neural Information Processing Systems*, pp. 2424–2432, 2016.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

Kleinberg, J., Liang, A., and Mullainathan, S. The theory is predictive, but is it complete?: An application to human perception of randomness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 125–126. ACM, 2017.

Mocanu, D. C., Mocanu, E., Stone, P., Nguyen, P. H., Gibescu, M., and Liotta, A. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9 (1):2383, 2018.

Mullainathan, S. and Spiess, J. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Noti, G., Levi, E., Kolumbus, Y., and Daniely, A. Behavior-based machine-learning: A hybrid approach for predicting human decision making. *arXiv preprint arXiv:1611.10228*, 2017.

Peysakhovich, A. and Naecker, J. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384, 2017.

Plonsky, O., Erev, I., Hazan, T., and Tennenholtz, M. Psychological forest: Predicting human behavior. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 656–662, 2017.

Plonsky, O., Apel, R., Erev, I., Ert, E., and Tennenholtz, M. When and how can social scientists add value to data scientists? A choice prediction competition for human decision making. Unpublished Manuscript, 2018.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., and Carter, E. C. Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*, 2019.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018.

Rosenfeld, A. and Kraus, S. Predicting human decision-making: From prediction to action. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(1): 1–150, 2018.

Russell, S. J. and Norvig, P. *Artificial intelligence: A modern approach*. Pearson Education Limited, 2016.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

Von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Yarkoni, T. and Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017.