AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

© 2024 American Psychological Association
ISSN: 2325-9965

2024, Vol. 11, No. 4, 619–632
https://doi.org/10.1037/dec0000242

Machine Learning for Modeling Human Decisions

Daniel Reichman[1], Joshua C. Peterson[2], and Thomas L. Griffiths[3, 4]

[1] Department of Computer Science, Worcester Polytechnic Institute
[2] Faculty of Computing and Data Sciences, Boston University
[3] Department of Psychology, Princeton University
[4] Department of Computer Science, Princeton University

The rapid development of machine learning has led to new opportunities for applying these methods to the study of human decision making. We highlight some of these opportunities and discuss some of the issues that arise when using machine learning to model the decisions people make. We first elaborate on the relationship between predicting decisions and explaining them, leveraging findings from computational learning theory to argue that, in some cases, the conversion of predictive models to interpretable ones with comparable accuracy is an intractable problem. We then identify an important bottleneck in using machine learning to study human cognition—data scarcity—and highlight active learning and optimal experimental design as a way to move forward. Finally, we touch on additional topics such as machine learning methods for combining multiple predictors arising from known theories and specific machine learning architectures that could prove useful for the study of judgment and decision making. In doing so, we point out connections to behavioral economics, computer science, cognitive science, and psychology.

Keywords: decision making, machine learning, active learning

The study of how people make decisions has a long and rich history (Baron, 2000; Hastie & Dawes, 2009) drawing on a diverse set of research methodologies such as normative models (Gilboa, 2009; Von Neumann & Morgenstern, 1947), behavioral experiments (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974), and computer simulations (Gigerenzer & Goldstein, 1996). Recently, a new research tool has become available for the study of decision making: combining machine learning (ML) models that excel in prediction with large data sets of human choices. This approach offers new ways to find predictive patterns that underlie choices, preferences, and decisions and is enjoying increasing popularity in the judgment and decision-making community, as well as more broadly in computer science (Rosenfeld & Kraus, 2018), economics (Kleinberg, Lakkaraju, et al., 2018; Mullainathan & Spiess, 2017), moral psychology (Agrawal et al., 2020; Awad et al., 2018), and neuroscience (Yamins et al., 2014). It appears that we are only beginning to discover the potential of using ML models and algorithms to uncover new aspects of human decision making.

The goal of this article is to point out theoretical questions arising from the study of human decisions and to demonstrate how ideas from the analysis of machine learning algorithms can shed light on these questions. In addition, we highlight a number of practical considerations based on our own experience that may be useful to scholars of decision making who may want to incorporate ML methods into their own research.

We mostly focus on *supervised learning*, wherein we are given a set of labeled examples and seek a model that predicts human behavior with low

Daniel Reichman https://orcid.org/0000-0003-0566-7528

Thomas L. Griffiths was supported by a grant from the NOMIS Foundation.

Daniel Reichman played a lead role in conceptualization, formal analysis, and writing–original draft. Joshua C. Peterson played supporting role in conceptualization, formal analysis, and writing–original draft. Thomas L. Griffiths played supporting role in conceptualization, formal analysis, and writing–original draft.

Correspondence concerning this article should be addressed to Daniel Reichman, Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, United States. Email: dreichman@wpi.edu

error on unseen examples (see Figure 1). This model is expressed as a function mapping inputs that correspond to choice problems to outputs that correspond to human decisions. For instance, given a training set consisting of people's choices between alternatives, we might seek a function that, given a pair of alternatives, predicts the probability of choosing each alternative for individuals in the population from which we are sampling. In the ideal scenario, this function would achieve low prediction error when evaluated on *unseen* decision problems (sets of alternatives) that were not in the "training set" from which the machine learning algorithm learns—a measure of how well the model generalizes. Often, the set of functions under consideration is defined by a set of parameters and the goal of the learning algorithm is to fit (*i.e.,* estimate) these parameters based on the training set to minimize the prediction error. A variety of approaches can be applied to this problem, corresponding to different choices of the set of possible functions. A familiar example for many researchers in the decision-making field is linear regression, where the set of possible functions comprises all linear combinations of features. Another example that is widely used in computer science but is less common in psychology research is deep neural networks (LeCun et al., 2015), which define a large class of nonlinear functions parameterized by the architecture and the weights assigned to simple units ("neurons") that are combined together to produce predictions.

A comprehensive survey of machine learning is beyond the scope of this work. The interested reader should consult machine learning texts such as Murphy (2012) or Hastie et al. (2009). More background on the theoretical foundations of machine learning can be found in Kearns and Vazirani (1994), Shalev-Shwartz and Ben-David (2014), and Hardt and Recht (2021).

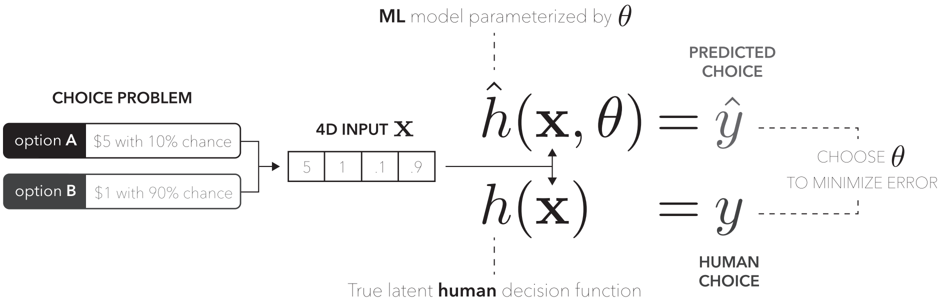
Can Machine Learning Be Used to Generate Novel Explanations of Human Decision Making?

Scientific inquiry into how people make decisions has a prediction component and an explanation component (Fudenberg et al., 2022; Plonsky et al., 2017). We want to predict which decision people will make when faced with a choice between alternatives, but we also want to explain *why* people prefer one alternative to the other. For example, prospect theory (Kahneman & Tversky, 1979) provides formulas to predict which of two lotteries are more likely to be chosen, but also explains why people favor some gambles more than others: people are more sensitive to losses than to gains, overestimate the likelihood of rare events, and may be influenced by a reference point when assessing the desirability of an outcome.

Machine learning models often excel at prediction. Prediction is typically measured by the error (e.g., mean squared error) on a test set that is disjoint from the data used to train the learning model. However, popular and effective machine learning methods, such as neural networks often result in opaque and complicated models that make it difficult to explain why a

Figure 1

Modeling Human Decisions as a Prediction Problem



Note. A choice problem is represented as a set of inputs that are provided to a machine learning algorithm. Human decision making is modeled as a function mapping these features to a decision (or a probability distribution over decisions). The parameters of the machine learning algorithm, θ, are chosen to minimize the difference between the estimated and true functions. ML = machine learning.

certain prediction is made. This is unsatisfying: while predictions are an important part of scientific theories, we also want to be able to come up with understandable and transparent explanations for the decisions people make.

From Prediction to Explanation

Can we leverage the superior power of machine learning models to improve our *understanding* of how and why people make decisions? This question has received attention recently both with respect to decision making (Fudenberg et al., 2022), as well as more generally for the social sciences (Hofman et al., 2021; Watts et al., 2018). For example, Fudenberg et al. (2022) have studied how to use machine learning models to test the completeness of theories: whether a current theoretical model captures all the variance in human behavior or is missing certain features. Those missing features can help us generate new explanations (Agrawal et al., 2020). Several studies have found that neural networks achieve lower prediction error when compared to theory-based computational models of human decision making (Ger et al., 2024; Peterson et al., 2021; Song et al., 2021). Comparing ML models to theory-based models is an underexplored avenue for gaining new insights into human behavior and has the potential to improve the accuracy of existing theories and reveal new features that explain people's decisions.

In one illustration of this approach, Peterson et al. (2021) proposed a method for understanding the theory about human behavior that a neural network may be implicitly discovering from a large data set. They used psychological theories of decision making to constrain the architectures of neural networks, making it possible to see how different assumptions influenced prediction performance. This is a way to evaluate the psychological theories—it makes it possible to find the best-performing model that satisfies the assumptions of the theory—but can also be used to provide insight into the behavior of neural networks in predicting decisions. If a theory-based model makes assumptions that parallel those discovered by an unconstrained neural network, then they will exhibit similar performance. Furthermore, the theory-based model will be able to reach this level of performance from less data. Peterson et al. (2021) used this approach to show that a modified version of prospect theory, wherein loss aversion and probability weighting become dependent on context, could perform as well as an unconstrained neural network in predicting human decisions.

The field of explainable machine learning is vast and covering it is beyond the scope of this article. Some examples of methods that are used to derive explainable machine learning models are local interpretable model-agnostic explanations (Ribeiro et al., 2016), partial dependence plots (Friedman, 2001), and Shapley Additive Explanations (Lundberg & Lee, 2017). Applying these methods to "black-box" models of human decision making has received, to our knowledge, little attention.

Theoretical Barriers to Explaining Human Decisions

Machine learning has been used to find new variables that predict and explain behavior. In Plonsky et al. (2017), the authors were able to extract novel "psychological features" to predict choice between gambles, and Fudenberg and Liang (2019) used decision trees to uncover new regularities explaining human play in two-player matrix games. Despite these successes, there are reasons to believe that prediction and explanation are qualitatively different problems (Shmueli, 2010). Here, we elaborate on one source of disparity that has received less attention, based on the computational complexity of extracting explainable models from black-box models.

Our argument is based on the difference between *proper* and *improper* learning (Kearns & Vazirani, 1994; Pitt & Valiant, 1988; Turán, 1994). To define what we mean by proper learning, we rely on the PAC (probably-approximate correct) learning framework (Valiant, 1984). The input to the learning algorithm is a training data set of independently identically distributed samples $(z_1, h(z_1)), \ldots (z_m, h(z_m))$ where h is an unknown function[1] from a class of functions $\mathcal{H}$ known the learner, and the z_i's belong to an n-dimensional space (e.g., the set of all Boolean vectors $\{0, 1\}^n$). Roughly speaking, in PAC-learning our goal is to output in polynomial time a function f from the family of functions $\mathcal{H}$ that with probability[2] at least

[1] Typically the range of all functions in $\mathcal{H}$ is $\{0, 1\}$. This setting is often referred to as PAC-learning binary classifiers.

[2] The probability is taken with respect to the distribution of the samples as well as the randomness of the learning algorithm.

$1-\delta$ results with a prediction error of at most ε where $\varepsilon, \delta \in (0, 1)$ are parameters that can be specified. Importantly, for a learning problem to be PAC-learnable, we insist the learning algorithm A is *proper*: It must output (given a data set of samples) a function that belongs to $\mathcal{H}$. We also require that the training algorithm A runs in time that is a polynomial function of the dimension of the input as well as $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and a size parameter s that measures the size of a representation of functions from $\mathcal{H}$ (Kearns & Vazirani, 1994).

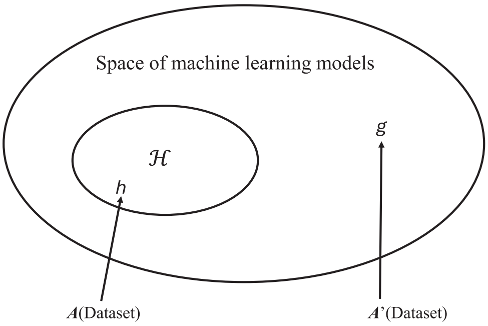
The generic way to design learning algorithms for PAC-learning is *empirical risk minimization*: find a function $f \in \mathcal{H}$ that perfectly fits the training data set[3]. Such a function will minimize the error of the chosen learning model on the training data set. That is, find $f \in \mathcal{H}$ such that for every $1 \leq i \leq m, f(z_i) = h(z_i)$. Assuming m is large enough and assuming some statistical properties of the hypothesis class $\mathcal{H}$, finding such an f efficiently guarantees that f will satisfy the PAC-learning requirements (Blumer et al., 1989).

For many choices of hypothesis classes, the computational complexity of finding a function f in $\mathcal{H}$ that perfectly fits the training data set is NP-hard and hence likely to be intractable (Pitt & Valiant, 1988; Turán, 1994). One way around such an intractability result is to relax the requirement that f belongs to $\mathcal{H}$ and return instead a function g that is not required to belong to $\mathcal{H}$. We still require that the prediction error of g is at most ε for an accuracy parameter ε and that the running time of the (improper) learning algorithm A' is polynomial in the parameters of the problem. We also insist that g can be computed efficiently (in polynomial time in the input dimensions). This weaker requirement on the learned function results in what is called an *improper* learning algorithm. When such an algorithm exists, we say that the problem is *predictable* (Turán, 1994). For an illustration of proper versus improper learning please see Figure 2.

Let us illustrate the concepts of proper and improper learning with an example that is related to learning rules that govern decisions between gambles. Consider the problem of learning the attractiveness of a gamble consisting of n monetary outcomes $x_1, \ldots, x_n \in \mathbb{R}$ occurring with probabilities $p_1, \ldots, p_n \in [0,1]$. We assume that the attractiveness of such a gamble takes a particular form, with $h_a(x_1 \ldots x_n, p_1 \ldots p_n) = \sum_{i=1}^{n} a_i x_i p_i$ for some vector $\bar{a} = (a_1, \ldots, a_n)$ of nonnegative real numbers (where larger values assigned to a gamble mean it is more likely to be chosen). The training data set consists of the parameters of pairs of gambles as well as the gambles chosen by the decision maker. In the *proper* learning case, we seek to output a function minimizing the prediction error from the family of all functions of the specific functional form $\{h_{\bar{a}}, \bar{a} \in \mathbb{R}^n_{\geq 0}\}$ where $\mathbb{R}_{\geq 0}$ is the set of all nonnegative real numbers. In the *improper* case, we want to find an *arbitrary* function g with low prediction error without requiring it to belong to the family $\{h_{\bar{a}}, \bar{a} \in \mathbb{R}^n_+\}$. For example, g could be a large neural network which was fit to the data using stochastic gradient descent.

Figure 2

An Illustration of Proper and Improper Learning and Their Relationship to Prediction and Explanation



Note. $\mathcal{H}$ is a family of machine learning models with a functional form that is useful for explaining decisions (e.g., halfspaces with coefficients in {0, 1}). A *proper* learning algorithm A takes a data set (such as list of pairs of gambles and human choices between the gambles) as input and outputs a machine learning model h in $\mathcal{H}$. An *improper* learning algorithm A' outputs a machine learning model g with low prediction error that is not guaranteed to belong to $\mathcal{H}$ (e.g., halfspace with real coefficients that are not all 0 or 1), and hence may fail to have a functional form that is useful for theoretical explanation of the data. In some cases, improper learning of a family of machine learning models $\mathcal{H}$ can be done efficiently, whereas proper learning of $\mathcal{H}$ is intractable. This represents a computational barrier for converting models with good predictive performance to models that are conducive to theoretical explanations.

[3] We assume there exists an f in $\mathcal{H}$ perfectly fitting the data. Similar reasoning can be applied when there is no function in $\mathcal{H}$ perfectly fitting that data and instead we look for $f \in \mathcal{H}$ that fits the data with minimum error Shalev-Shwartz and Ben-David (2014).

As another illustration, consider the problem of *intertemporal choice* (Berns et al., 2007). We follow closely the exposition in Chase and Prasad (2019): our goal is to predict which of two payoff vectors $(x_1, \ldots, x_T)$ and $(y_1, \ldots, y_T)$ both in $\mathbb{R}^T$ a decision maker will choose, where there are $T > 1$ discrete time slots and x_i, y_i are the payoffs received at time $1 \leq i \leq T$. One way to model the preference between payoffs is to assume that $(x_1, \ldots, x_T)$ is preferable to $(y_1, \ldots, y_T)$ if only if $\sum_{i=1}^{T} D(i)x_i \geq \sum_{i=1}^{T} D(i)y_i$ where $D: \{1, \ldots T\} \rightarrow (0, 1)$ belongs to a family of nonincreasing utility functions. For example, $D(i) = \delta^i$ represents *exponential discounting* (where $\delta \in (0, 1)$) and $D(i) = \beta\gamma^{i-1}$ with $\beta, \gamma \in (0, 1)$ represents *Quasi-hyperbolic discounting* (Laibson, 1997) capturing a "presence bias." In proper learning, we seek to output a hypothesis from the family of utility functions (e.g., find a $\delta \in (0, 1)$ minimizing the prediction error for exponential discounting). In improper learning, we seek an efficiently computable model $f: \mathbb{R}^T \rightarrow \mathbb{R}$ such that $(x_1, \ldots, x_T)$ is preferable to $(y_1, \ldots, y_T)$ if only iff $f(x_1, \ldots, x_T) \geq f(y_1, \ldots, y_T)$. As in the previous example, f does not have to be of the form of any particular utility function and can be a complicated black-box model.

Proper learning is related to explanation in the following way: One form of "explanation" could be to require that the learned model predicting human decision making will have a specific functional form such as expected utility. Explanations could be achieved by learning a function with a specific form, which corresponds to proper learning. However, it is well known that for several learning problems improper learning is tractable (can be done in polynomial time), yet proper learning is NP-hard and hence likely to be computationally intractable (Kearns & Vazirani, 1994; Shalev-Shwartz & Ben-David, 2014; Turán, 1994). Put differently, there are cases where improper learning might require significantly less computational resources than proper learning. It follows that even if we can find a predictor with small prediction error efficiently, it could be intractable to deduce from it an explainable or interpretable model of a desired functional form. Our main argument is that in such cases there will be a computational divide between prediction and explanation: Finding an explainable machine learning model from a hypothesis class is intractable, even though we can efficiently find a learning model with arbitrarily small prediction error.

As a concrete example where converting a model to a model with superior explainable properties is computationally intractable, consider a set of learning models called *halfspaces*. A 0–1 halfspace is parameterized by a Boolean vector $\bar{a} = (a_1 \ldots a_d) \in \{0, 1\}^d$ and an integer b. This defines a function that classifies its input space, with halfspace $h_{\bar{a},b}$ applied to a Boolean input vector $(z_1, \ldots, z_\mathrm{d})$ resulting in an output of 1 if $\sum_{i=1}^{d} a_i z_i \geq b$ and 0 otherwise. It is known that properly learning this hypothesis class is NP-hard (Pitt & Valiant, 1988). However, this learning problem is predictable: We can efficiently find a predictor that will have arbitrary small prediction error. To see why, suppose we now consider the hypothesis class $\mathcal{H}'$ which consists of all halfspaces parameterized by a *real* vector $\bar{c} = (c_1 \ldots c_d) \in \mathbb{R}^d$ and a threshold $d \in \mathbb{R}$. Similarly $h_{\bar{c},d}((z_1, \ldots, z_d)) = 1$ if $\sum_{i=1}^{d} c_i z_i \geq d$ and 0 otherwise. This learning problem can be solved efficiently (Shalev-Shwartz & Ben-David, 2014) as there are polynomial time algorithms (such as linear programming) that can find a halfspace $h_{\bar{c}',d'}$ that perfectly fit a training data set (assuming there is a predictor in $\mathcal{H}'$ achieving zero error)[4]. Hence, the learning problem is predictable, since we can efficiently solve the empirical risk minimization problem with respect to $\mathcal{H}'$. The catch is that the halfspace $h_{\bar{c}',d'}$ we get in this way does not belong to the original hypothesis class $\mathcal{H}$: We have no guarantees that the (real) parameters of the found halfspace are in $\{0, 1\}$ Observe that intuitively $\mathcal{H}$ is more interpretable than the set $\mathcal{H}'$ of all halfspaces with real parameters and hence might be more useful in explaining why a classification was made: It marks the "important features" that influence the classification (those with Coefficient 1).

To illustrate the distinction between arbitrary as opposed to binary halfspaces in studying decisions, consider the following scenario. Suppose we use a halfspace as a simple model for multiattribute decision making (Payne et al., 1993), predicting which binary attributes influence choices people make when considering options with several attributes (e.g., choosing a car based on attributes such as whether it is electric, has white color and

[4] Here, we also rely on the fact that the sample complexity of learning halfspaces is polynomial: We ensure arbitrarily small prediction error ε with probability at least $1-\delta$ by taking the number of samples m to be a polynomial in n, $1/\varepsilon$, $1/\delta$

price lower than 1000 Euros). To simplify the exposition, we focus on the case of a single option that can be chosen or not, where an output of 1 marks choosing the option. A 0–1 halfspace offers a simple interpretation "selecting" all features that influence the decision maker. The meaning of coefficients in arbitrary halfspaces which might be arbitrary real numbers could be more complicated to interpret.

An important property of explainable models is succinctness. Even classifiers that are perceived as more useful in deriving explanations, such as decision trees, become incomprehensible as their size grows: A decision tree with hundreds (or more) nodes is hard to interpret or understand. Many of the machine learning models (such as neural networks) that excel in prediction are over-parameterized, meaning the number of parameters far exceeds the size of the training data set (Hardt & Recht, 2021). When such large models are used on scientific data, their size can make it a challenge to gain new explanations from their superior predictive performance. It is computationally intractable (NP-hard), to decide whether a given data set can be fit with a classifier with a fixed size limit (Blum & Rivest, 1992; Goel et al., 2021). This creates another theoretical divide between prediction and explanation: While we may be able to obtain a large unconstrained ML model with small prediction error, it may be computationally infeasible to extract from it a compact model of potential use to theoretical understanding.

Another obstacle to obtaining explanations from black-box models such as neural networks is that it is challenging to certify whether they have "natural" properties (Lin & Vitter, 1991) we expect from a model that explains decision making. For example, it is shown in Lin and Vitter (1991) that for certain architectures even simple questions such as whether a given network is not identically zero is intractable. As another example, it is notoriously hard to certify that a neural network computes a monotone function (Sivaraman et al., 2020) whereas monotonicity (or lack of thereof) is an important property in understanding models of human decision making.

Finally, it should be remembered that decision making is a highly complex process (Livnat & Pippenger, 2008) that can be influenced by numerous features including a stochastic component that cannot be modeled by deterministic rules (Davis-Stober & Brown, 2011). This complexity necessarily limits the predictive accuracy of compact, theoretically appealing models. Hence, it is to be expected that a trade-off between explainability and accuracy (Lakkaraju et al., 2016) in modeling how people make decisions is unavoidable. In particular, as we obtain more data about human decisions we are going to be able to infer more complex models from it (Peterson et al., 2021). This is a direct consequence of a fundamental principle of statistics known as the *bias–variance trade-off* (Geman et al., 1992). With limited data simpler models result in better generalization, because the greatest risk to generalization is overfitting the data and hence producing predictions that vary significantly across different data sets. As the amount of data increases, this kind of variance is less of a risk and more complex models are the best path toward generalization because they are capable of producing an unbiased estimate of the underlying function.[5] Psychology has lived in the limited data regime since its inception and has hence focused on simple models. As we move to larger and larger data sets, we are going to need to become more comfortable with complexity (Griffiths, 2015).

We stress that the obstacles outlined in this section do not mean that leveraging machine learning as a way of generating new explanations of human decisions should be abandoned. Sometimes explainable learning models of human choice can be learnt efficiently: For example, exponential and Quasi-hyperbolic discounting methods admit efficient learning algorithms (Chase & Prasad, 2019). However, the results in this section do point to the conclusion that obtaining explanations from models that make superior predictions is likely to be more computationally demanding compared to achieving superior predictions alone.

Active Learning and Data Collection

Empirical studies of human decision making typically include no more than a few dozen choice

[5] It is worth noting that the term "bias" used here corresponds to a formal statistical notion—that the average of the estimated functions across data sets is close to the true function—which differs from the broader notions of bias used in the study of judgment and decision making and in psychology more generally.

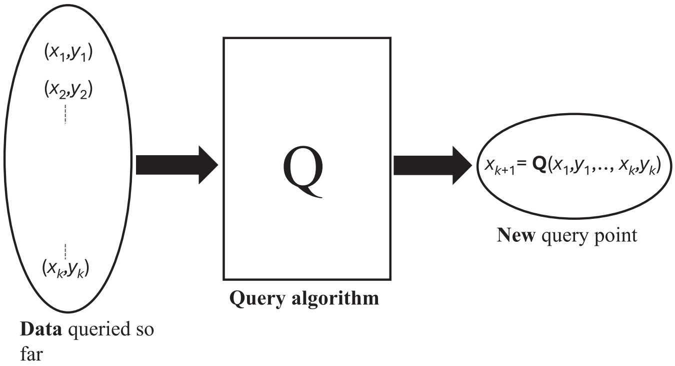
problems, where each problem is a choice between several gambles with varying parameters (monetary rewards, probabilities of gains and losses). Machine learning models such as neural networks are typically data-hungry and need data sets with thousands of data points (or more) to improve upon theoretical models. The reliance on increasingly large sets of items is also the trend in recent studies (Erev et al., 2017; Peterson et al., 2021). With these developments, it is no longer feasible to manually construct experimental stimuli and it is necessary to resort to algorithms that generate these stimuli automatically. Designing theoretically informed algorithms for solving this problem that are simple to implement is an important goal. Such algorithms are likely to be useful in other domains which rely on large data sets such as categorization (Battleday et al., 2020).

Supervised learning assumes that the learner has little control over the datapoints it examines: The learner is restricted to randomly choosing datapoints and then applying a machine learning algorithm on the resulting data set to obtain an ML model with low prediction error on unseen data. In contrast, *active learning* (Dasgupta, 2011; Zhu et al., 2018) refers to a scenario in which it is possible to choose data points *adaptively* and arbitrary datapoints can be queried in the course of constructing a machine learning model (Figure 3). Active learning can lead to significant decrease in the sample complexity required to output a machine learning model such as halfspace achieving a desirable accuracy level: For an excellent illustration see Zhu et al. (2018). However, active learning algorithms (compared to gradient based learning methods of neural networks) are often more complicated to implement efficiently and applying them to large data sets has been a challenge. In behavioral experiments, the adaptive choice of which queries to present to participants in psychological experiments is also referred as *optimal experimental design* (Cavagnaro et al., 2010; Myung et al., 2013). The idea is that adaptively selecting queries based on the current performance (i.e., prediction error) of the machine learning model could reduce the amount of data needed to attain a desired prediction error. Such techniques are important in behavioral research, where recruiting participants is very costly. The large number of participants needed to generate data sets for training ML models is a major bottleneck in using machine learning in behavioral research. Therefore, active learning has the potential to make the of ML for behavioral experiments more viable.

To return to our example of predicting people's choices between pairs of gambles, the active

Figure 3

Illustration of an Iteration of the Active Leaning Pipeline



Note. The query algorithm has access to $k > 0$ previously queried datapoints, where x_{i} is a feature vector (e.g., numerical quantities of pairs gambles) and y_{i} is the corresponding label (e.g., whether the first or second gamble was chosen by the human subject). Based on these k data points, the algorithm outputs a new query point x_{k+1} (payoffs and probabilities of pairs of gambles) and queries the decisions made by human participants on this gamble. Once an appropriate stopping criterion is reached, the algorithm outputs an ML model based on all the data points queried. ML = machine learning.

learning problem becomes one of finding the pairs that will be most informative for training a model. The researcher has the choice of which gambles to present depending on the error of the current classifier, as well as properties of the person participating in the experiment (demographic features, past choices). The question is then how to choose gambles in an informed and adaptive way to make best use of the resulting samples.

Active learning methods often work by looking for a datapoint that leads to the maximum increase in expected information gain with respect to the model we are trying to learn (Foster et al., 2019). One method is to choose datapoints that current models are maximally uncertain about, For example, one may choose a pair of gambles where the currently constructed model gives probability roughly $\frac{1}{2}$ to each choice, meaning it cannot do much better than chance when deciding which gamble will be chosen by the typical participant. Such active learning methods have been used for reward learning in human–robot interactions (Reddy et al., 2020). Applying these methods to support the rapid development of new theories of human decision making is an interesting direction for future research.

The use of adaptive methods in discriminating between models of risky choice was examined in Cavagnaro et al. (2013) using extensive simulations. A different active learning algorithm for learning risky choice along with theoretical study of the guarantees of the algorithm such as sample complexity is described in Echenique and Prasad (2020). Applying methods of this kind to construct machine learning models of human decision making (based on human data) is a promising direction, as it could lead to more compact data sets that achieve comparable accuracy to state of the art models and data sets.

Active learning can also be used to obtain interpretable models. Finite automata are used to model and understand boundedly rational agents (Wilson, 2014). A classic result due to Angluin (1988) is that a finite automaton can be learned efficiently by using a sophisticated active leaning algorithm. Applying this algorithm to construct automata that model human decision making could lead to new theoretical findings.

Up to this point, we have mostly focused on an experimental paradigm that asks people to choose between gambles based on their numerical parameters. In many situations, decisions are likely to be based on additional sources of information such as text (Apel et al., 2022) as well as images (Hilgard et al., 2021). Predicting and understanding how people make decisions based on visual stimuli has received attention lately (Bhatia & Stewart, 2018; Trueblood et al., 2021). The recent advancements in generative artificial intelligence open the door to automatically constructing more sophisticated stimuli (Peterson et al., 2022) that can be used to study how people make decisions for visual, textual, and multimodal stimuli.

Beyond the considerations of stimulus selection, there is also the parallel problem of obtaining high-quality human responses to these stimuli. Erev et al. (2017) emphasized quality by obtaining large numbers of responses from different raters for each choice problem stimulus, whereas Peterson et al. (2021) emphasize quantity instead by collecting fewer responses but for more choice problems overall. While both presumably boost the predictive power of ML algorithms, estimating the optimal trade-off between the number of stimuli and the number of responses per stimulus is still an open problem.

Despite the potential utility of active learning as a strategy for dealing with the data needs of ML algorithms, we urge caution in using these methods. Actively selecting datapoints to discriminate between current models can result in a systematically skewed sample, running the risk of focusing on an unrepresentative subset of human decisions and being overly influenced by the set of models under consideration. This can make the resulting data sets difficult to use as benchmarks for evaluating new models. While we see active learning as a good tool for deploying ML models more widely in the short term, in the long term, we should work to complement these efforts with large-scale model-agnostic samples. For further discussion of the potential of these different approaches see Almaatouq et al. (2024).

Further Directions

Combining Multiple Theories

There are often many competing theories within a given domain of decision making. For instance, He et al. (2022) identified at least 62 models of risky choice. Rather than assuming that a single theory offers a more accurate characterization of people's decisions than the others, some researchers have considered the idea that individual theories could be *combined* to build new, better theories. For example, Davis-Stober and Brown (2011) proposed a method to find convex combinations of the predictions of multiple models that better predict behavior than individual models, and Erev et al. (2017) recently proposed best estimate and sampling tools, a successful model of risky choice that utilizes a linear combination of different value functions (*i.e.*, formulas for assigning value to a gamble), yielding highly competitive predictions.

Extending these approaches and taking inspiration from *mixture-of-experts* models in machine learning, Peterson et al. (2021) proposed a method to infer predictive mixtures at the more granular level of individual choice problems. This additional flexibility allows for the possibility that decision makers might think differently depending on the particular decision problem they are facing. This model, called mixture of theories, defines the value $V(A)$ of a gamble A as a weighted sum of classic utility and probability weighting functions (u_j and π_k; Kahneman & Tversky, 1979):

$$V(A) = \sum_{i \in A} \left[\sum_j \omega_j u_j(x_i) \right] \left[\sum_k \omega_k \pi_k(p_i) \right]. \quad (1)$$

Importantly, mixture weights ω_j and ω_j (where $\omega_j = 1 - \omega_k$) are not fixed but instead vary as the outputs of a neural network that takes as input an entire decision problem. This neural network is therefore learning to predict which mixture weights best apply to each decision problem. Varying weights in this particular model amounts to dynamically modulating the influence of classic effects such as loss aversion and probability weighting (Kahneman & Tversky, 1979), as mentioned in an earlier section. While most of the above approaches happen to make use of value functions as mixture components, the principle of mixing decision strategies can be applied to nearly any theoretical tradition (*e.g.*, heuristics).

An influential method in machine learning for making productive use of multiple models is *boosting* (Freund & Schapire, 1997)[6]. This method combines several binary classifiers to create a single model with lower prediction error using majority vote (similar approaches work for real valued predictors by taking the mean of the different predictions). Boosting can reduce the error of single models (called "weak learners") that have limited ability to obtain nearly minimal prediction error. This approach can be useful when applied to psychological theories that are derived from laboratory experiments, which often excel in theoretical vigor but can have low predictive accuracy. Plonsky et al. (2017) applied a boosting technique called *random forests* to a choice prediction task, demonstrating improved predictive performance when compared to recently developed theoretical models such as Best Estimate and Sampling Tools (Erev et al., 2017). In contrast to neural networks which require large data sets with tens of thousands of data points (or more), these results in (Plonsky et al., 2017) suggest that boosting could be a useful tool for the study of human decisions on small data sets with a few hundreds (or less) of choices between gambles. For example, the winners of the 2018 International Choice Prediction Competition applied exactly this strategy (Plonsky et al., 2019). Applying boosting to obtain new theories of human decision making as well as psychological theories in other domains is an exciting future direction.

Using Theoretical Machine Learning in Theories of Human Decision Making

So far, we have discussed how machine learning models are instrumental in predicting and understanding decisions. Another direction that has received significantly less attention is leveraging findings from theoretical machine learning in building new theories of judgement and choice. Recently, Haghtalab et al. (2021) have used theoretical machine learning to provide new explanatory models for the emergence of belief polarization. Formal principles arising from stochastic optimization (which is used to fit machine learning models) have been used to explain how people integrate appraisals from the environment into their affective states, which can influence decisions (Bennett et al., 2022). We believe that this is the tip of the iceberg—theoretical machine learning has much to offer to theories of decision-making and human cognition more generally.

[6] This kind of "boosting" is not in any way related to the usage of the term in the decision making literature on improving the decision people make.

Theoretical analysis of combinatorial parameters related to the sample complexity of choice models such as Vapnik–Chervonenkis (VC) dimension[7] or other quantities related to efficiency of learning for normative models of decision making is uncommon. One exception is the work of Chase and Prasad (2019), who analyzed the VC dimension of time-dependent choice models such exponential, hyperbolic and Quasi-hyperbolic discounting and proved that these models can be learnt efficiently. Other examples have considered the learnability of preference in the revealed preference setting (Balcan et al., 2014; Basu & Echenique, 2020; Beigman & Vohra, 2006). Additional investigation of the sample and time complexity of choice models can be useful to theoretical developments of normative models of decision making, especially if the limitations of bounded rational learners are considered (Haghtalab et al., 2021). It could also lead to useful learning algorithms that can be applied to data sets of human choice.

Which Learning Model to Choose

There are a vast array of machine learning models to choose from when fitting behavioral data. Even when restricted to neural networks there are many options—these models can vary in the number of neurons, the depth of the network, the activation function, and parameters of the learning algorithms such as learning rate. While the use of ML in the study of human decision making and psychological data more generally is still on its infancy, work in this area is quickly expanding as researchers begin to explore different neural network architectures such as convolutional neural networks, transformers, and recurrent neural networks (Dezfouli et al., 2019; Tuli et al., 2021).

Achieving state of the art performance in tasks such as image recognition and text processing often requires large neural networks with millions of neurons and dozens of layers. However, it appears that much smaller networks are sufficient to improve upon the prediction error of theoretical models of human decision making. For example, Peterson et al. (2021) used a network with just 32 neurons and two hidden layers. Network pruning, leading to an architecture with fewer connections, is one method for reducing the size and prediction error of such models (Bourgin et al., 2019).

One model that has proven useful in the study of sequential decision making by people is recurrent neural networks (RNNs). These neural network models have been used to model text and language generation and are useful for sequential prediction tasks that depend on a sequence of inputs with temporal structure (Elman, 1990). Energy-based recurrent models inspired by Hopfield networks have been used to model human decision (Glöckner et al., 2014). More recently, several studies have shown that RNNs achieve lower prediction error when compared with normative models of predictive choice such as reinforcement learning (Ger et al., 2024; Ji-An et al., 2023; Song et al., 2021). Furthermore, the superior predictive efficacy of recurrent networks has been used to learn parameters in a reinforcement learning task, resulting with an interpretable model that achieves superior accuracy to comparable models that are constructed based on theoretical principles alone (Ger et al., 2024). RNNs could find uses beyond sequential decision-making and reinforcement learning. For example, RNNs could be useful for the standard paradigm of studying risky choices where blocks of decisions between gambles are presented to participants: The use of recurrent architectures could shed light on how choices people face early in the experiment can influence later decision. Further study of RNNs in predicting and understanding how people handle sequential decisions over time is an interesting direction for future research.

To summarize, there are many possible choices of neural network architectures when fitting behavioral data, and current research suggest that networks with a few hundred of neurons that can be trained in a reasonable amount of time can already point out to interesting predictive patters that are not captured by theoretical models. One future direction could be to systematically compare the performance of different architectures on current data sets of human risky choice pointing to which architectures excel in such tasks.

Ethical Considerations

While the use of machine learning models and generative AI has the potential to advance our

[7] VC dimension is a combinatorial parameter that characterizes the sample complexity of learning problems in the PAC learning model Vapnik (1999). For more details please see Shalev-Shwartz and Ben-David (2014)

scientific understanding, it can also be used to develop methods to manipulate people to make decisions that are not in their best interests (Dezfouli et al., 2020). Researchers should be aware of these dangers and take measures to avoid the data they collect, models that they build, and procedures for relating realistic stimuli to human decisions being used in unethical ways. It should be kept in mind that questions regarding decision making can reveal information about thought process of participants that they might not be interested to share (Hilgard, 2021). Therefore, efforts should be made to make people fully aware about the information they are providing while participating in the experiment. Researchers should also be aware that even when they choose not to make public a sensitive data set, the training data might be reconstructed from machine learning models (Carlini et al., 2021, 2022) and could lead to privacy issues. Developing formal measures of the potential risks of data sets documenting human choice and method to mitigate these risks is consequently of great interest. Researchers should also consider potential biases in their data sets that may fail to represent appropriately certain groups. This could pose risks to the validity of the conclusions drawn about human behavior. Special care should be taken when using machine learning models as decision-support tools as these could lead to discrimination. For more on this issue and potential solutions see Kleinberg, Ludwig, et al. (2018).

Conclusion

Machine learning offers many new opportunities for the study of human decision making. We have outlined some of the theoretical and practical issues that arise when applying ML to behavioral data sets of choice behavior, such as the challenges of extracting explanations from predictive models, employing active learning methods, and choosing useful model architectures. We further explored how findings from theoretical machine learning could shed light on some of these questions and considerations. We believe there are more connections to be made between theory development for human decision making and computational learning theory, creating significant opportunities for future research. As we begin to work with larger and larger data sets of human decisions, ML methods are going to become increasingly important as a supplement to traditional theory-building, equipping us with new tools to make sense of complex behaviors that might otherwise be hard for human scientists to discover alone.

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, *117*(16), 8825–8835. https://doi.org/10.1073/pnas.1915841117

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*, Article e33. https://doi.org/10.1017/S0140525X22002874

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319–342. https://doi.org/10.1023/A:1022821128753

Apel, R., Erev, I., Reichart, R., & Tennenholtz, M. (2022). Predicting decisions in language based persuasion games. *Journal of Artificial Intelligence Research*, *73*, 1025–1091. https://doi.org/10.1613/jair.1.13510

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Balcan, M.-F., Daniely, A., Mehta, R., Urner, R., & Vazirani, V. V. (2014). Learning economic parameters from revealed preferences. In T-Y. Liu, Q. Qi, & Y. Ye (Eds.), *Web and internet economics: 10th international conference, wine 2014, Beijing, China, December 14–17, 2014* (Vol. 10, pp. 338–353). Springer International Publishing.

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

Basu, P., & Echenique, F. (2020). On the falsifiability and learnability of decision theories. *Theoretical Economics*, *15*(4), 1279–1305. https://doi.org/10.3982/TE3438

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), Article 5418. https://doi.org/10.1038/s41467-020-18946-z

Beigman, E., & Vohra, R. (2006). Learning from revealed preference. In J. Feigenbaum, J. Chuang, & M. D. Pennock (Eds.), *Proceedings of the 7th ACM conference on electronic commerce* (pp. 36–42). Association for Computing Machinery.

Bennett, D., Davidson, G., & Niv, Y. (2022). A model of mood as integrated advantage. *Psychological Review*, *129*(3), 513–541. https://doi.org/10.1037/rev0000294

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice–toward an integrative framework. *Trends in Cognitive Sciences*, *11*(11), 482–488. https://doi.org/10.1016/j.tics.2007.08.011

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88. https://doi.org/10.1016/j.cognition.2018.05.025

Blum, A. L., & Rivest, R. L. (1992). Training a 3-node neural network is NP-complete. *Neural Networks*, *5*(1), 117–127. https://doi.org/10.1016/S0893-6080(05)80010-3

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, *36*(4), 929–965. https://doi.org/10.1145/76359.76371

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. In K. Chaudhuri & R. Salakhutdinov (Eds.), *International conference on machine learning* (pp. 5133–5141). Proceedings of Machine Learning Research.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022). Membership inference attacks from first principles. In T. Holz & T. Ristenpart (Eds.), *2022 IEEE symposium on security and privacy (SP)* (pp. 1897–1914). Institute of Electrical and Electronics Engineers.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., & Oprea, A. (2021). Extracting training data from large language models. In M. Bailey & R. Greenstadt (Eds.), *USENIX security symposium* (Vol. 6). USENIX Association.

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., & Pitt, M. A. (2013). Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management Science*, *59*(2), 358–375. https://doi.org/10.1287/mnsc.1120.1558

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887–905. https://doi.org/10.1162/neco.2009.02-09-959

Chase, Z., & Prasad, S. (2019). *Learning time dependent choice* [Conference session]. 10th Innovations in Theoretical Computer Science Conference (ITCS 2019), University of California San Diego in San Diego, California.

Dasgupta, S. (2011). Two faces of active learning. *Theoretical Computer Science*, *412*(19), 1767–1781. https://doi.org/10.1016/j.tcs.2010.12.054

Davis-Stober, C. P., & Brown, N. (2011). A shift in strategy or "error"? strategy classification over multiple stochastic specifications. *Judgment and Decision Making*, *6*(8), 800–813. https://doi.org/10.1017/S1930297500004228

Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P., & Balleine, B. W. (2019). Models that learn how humans learn: The case of decision-making and its disorders. *PLOS Computational Biology*, *15*(6), Article e1006903. https://doi.org/10.1371/journal.pcbi.1006903

Dezfouli, A., Nock, R., & Dayan, P. (2020). Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences*, *117*(46), 29221–29228. https://doi.org/10.1073/pnas.2016921117

Echenique, F., & Prasad, S. (2020). Incentive compatible active learning. In T. Vidick (Ed.), *11th innovations in theoretical computer science conference [ITCS 2020]* (pp. 1–20). Dagstuhl Publishing.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369–409. https://doi.org/10.1037/rev0000062

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., & Goodman, N. (2019). Variational Bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, *32*, 14036–14047.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/101320345

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, *130*(4), 956–990. https://doi.org/10.1086/718371

Fudenberg, D., & Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*, *109*(12), 4112–4141. https://doi.org/10.1257/aer.20180654

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Ger, Y., Nachmani, E., Wolf, L., & Shahar, N. (2024). Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior. *PLOS Computational Biology*, *20*(1), Article e1011678. https://doi.org/10.1371/journal.pcbi.1011678

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gilboa, I. (2009). *Theory of decision under uncertainty* (Vol. 45). Cambridge University Press.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, *133*(3), 641–666. https://doi.org/10.1016/j.cognition.2014.08.017

Goel, S., Klivans, A., Manurangsi, P., & Reichman, D. (2021). Tight hardness results for training depth-2 ReLU networks. In J. Lee (Ed.), *12th innovations in theoretical computer science conference [ITCS 2021]* (pp. 22:1–22:12). Dagstuhl Publishing.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23. https://doi.org/10.1016/j.cognition.2014.11.026

Haghtalab, N., Jackson, M. O., & Procaccia, A. D. (2021). Belief polarization in a complex world: A learning theory perspective. *Proceedings of the National Academy of Sciences*, *118*(19), Article e2010144118. https://doi.org/10.1073/pnas.2010144118

Hardt, M., & Recht, B. (2021). *Patterns, predictions, and actions: A story about machine learning*. arXiv preprint arXiv:2102.05242.

Hastie, R., & Dawes, R. M. (2009). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

He, L., Zhao, W. J., & Bhatia, S. (2022). An ontology of decision models. *Psychological Review*, *129*(1), 49–72. https://doi.org/10.1037/rev0000231

Hilgard, A. S. (2021). *Machine learning for humans: Building models that adapt to behavior* (Unpublished doctoral dissertation). Harvard University.

Hilgard, S., Rosenfeld, N., Banaji, M. R., Cao, J., & Parkes, D. (2021). Learning representations by humans, for humans. In M. Melia & Z. Tong (Eds.), *International conference on machine learning* (pp. 4227–4238). Proceedings of Machine Learning Research.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., & Vespignani, A. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Ji-An, L., Benna, M. K., & Mattar, M. G. (2023). *Automatic discovery of cognitive strategies with tiny recurrent neural networks*. bioRxiv, 2023–04.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Kearns, M. J., & Vazirani, U. (1994). *An introduction to computational learning theory*. MIT Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *Aea Papers and Proceedings*, *108*, 22–27. https://www.aeaweb.org/articles?id=10.1257/pandp.20181018

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478. https://doi.org/10.1162/003355397555253

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In A. Smola & C. Aggarwal (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675–1684). Association for Computing Machinery.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lin, J.-H., & Vitter, J. S. (1991). Complexity results on learning by neural nets. *Machine Learning*, *6*, 211–230. https://doi.org/10.1007/BF00114777

Livnat, A., & Pippenger, N. (2008). Systematic mistakes are likely in bounded optimal decision-making systems. *Journal of Theoretical Biology*, *250*(3), 410–423. https://doi.org/10.1016/j.jtbi.2007.09.044

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* [Conference session]. Advances in Neural Information Processing Systems.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3–4), 53–67. https://doi.org/10.1016/j.jmp.2013.05.005

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

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. https://doi.org/10.1126/science.abe2629

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), Article e2115228119. https://doi.org/10.1073/pnas.2115228119

Pitt, L., & Valiant, L. G. (1988). Computational limitations on learning from examples. *Journal of the ACM*, *35*(4), 965–984. https://doi.org/10.1145/48014.63140

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

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). *Psychological forest: Predicting human behavior* [Conference session]. Proceedings of the AAAI Conference on Artificial Intelligence.

Reddy, S., Dragan, A., Levine, S., Legg, S., & Leike, J. (2020). Learning human objectives by evaluating hypothetical behavior. In H. Daume III & A. Singh (Eds.), *International conference on machine learning* (pp. 8020–8029). Proceedings of Machine Learning Research.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “why should I trust you?” Explaining the predictions of any classifier. In A. Smola & C. Aggarwal (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery.

Rosenfeld, A., & Kraus, S. (2018). Predicting human decision-making: From prediction to action. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *12*(1), 1–150. https://doi.org/10.2200/S00820ED1V01Y201712AIM036

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Sivaraman, A., Farnadi, G., Millstein, T., & Van den Broeck, G. (2020). *Counterexample-guided learning of monotonic neural networks* [Conference session]. Advances in Neural Information Processing Systems.

Song, M., Niv, Y., & Cai, M. (2021). *Using recurrent neural networks to understand human reward learning* [Conference session]. Proceedings of the Annual Meeting of the Cognitive Science Society.

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O’Daniels, P., & Holmes, W. R. (2021). Disentangling prevalence induced biases in medical image decision-making. *Cognition*, *212*, Article 104713. https://doi.org/10.1016/j.cognition.2021.104713

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. (2021). *Are convolutional neural networks or transformers more like human vision?* [Conference session]. Proceedings of the Annual Meeting of the Cognitive Science Society.

Turán, G. (1994). Computational learning theory and neural networks: A survey of selected topics. In V. Roychowdhury, K. Y. Siu, & A. Orlitsky (Eds.), *Theoretical advances in neural computation and learning* (pp. 243–293). Springer.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142. https://doi.org/10.1145/1968.1972

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer Science+Business Media.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd rev.).

Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., Grubesic, A., Hofman, J. M., Rohrer, J. M., & Salganik, M. (2018). *Explanation, prediction, and causality: Three sides of the same coin?* https://osf.io/preprints/osf/u6vz5

Wilson, A.(2014). Bounded memory and biases in information processing. *Econometrica*, *82*(6), 2257–2294. https://doi.org/10.3982/ECTA12188

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). *An overview of machine teaching*. arXiv preprint arXiv:1801.05927.

Received May 30, 2023
Revision received March 14, 2024
Accepted April 30, 2024 ■