Using Machine Learning to Generate, Clarify, and Improve Economic Models*

Annie Liang†

February 15, 2026

Abstract

This article examines how machine learning can contribute to economic models that explain behavior, rather than merely predicting outcomes. First, machine learning predictions can be used as benchmarks for economic models, for instance to quantify how far they are from the predictive limit. Second, algorithms can adversarially probe economic models to identify cases in which the theoretical predictions fail. Third, "hybrid" models can combine interpretable economic structure with flexible learning methods to leverage the strengths of both. Finally, large language models introduce qualitatively new possibilities, from simulating human responses to generating novel hypotheses. Throughout, I emphasize both promise and limits: while machine learning can uncover patterns that standard models miss, translating these algorithmic insights into interpretable and portable economic understanding typically still requires human judgment.

1 Introduction

Machine learning algorithms can now outperform classic economic models in predicting outcomes ranging from bargaining agreements (Camerer, Nave, and Smith, 2019) to choice under uncertainty (Ellis, Kariv, and Ozbay, 2023a) to future jobs and wages

*I am grateful to David Romer and five anonymous referees for helpful feedback, and to Drew Fudenberg, Sendhil Mullainathan, Isaiah Andrews, Wayne Gao, Sukjin Han, Jon Kleinberg, and Lihua Lei for many conversations that shaped my views on this topic.

†Department of Economics, Northwestern University.

(Vafa, Palikot, Du, Kanodia, Athey, and Blei, 2024b). Yet this predictive accuracy comes at a cost: most machine learning algorithms function as black boxes, offering little insight into *why* these outcomes occur.

This article asks whether machine learning can guide the development of new economic theories that explain behavior rather than merely predict it. Economic analysis often aims to guide interventions, policy, and institutional design despite limited data. We thus value models that identify mechanisms and causal pathways that can be acted upon: a framework that forecasts which countries will escape poverty allows us to prepare for the future, while a framework that clarifies how development occurs may allow us to change it. We also value models that generalize across environments, as this allows insights drawn from one setting to inform decisions in others.

This paper examines the current relationship between economic modeling and machine learning, and explores potential complementarities between these approaches. At the present, off-the-shelf application of machine learning algorithms is unlikely to yield useful new economic theories. But recent work shows that machine learning methods can nevertheless help us improve, understand, and build upon existing models. These approaches range from adversarially training algorithms to expose the limits of existing models, to imposing economic theory as a constraint on algorithmic search. Recent advances in large language models both complement these strategies and open entirely new research directions.

The plan of the paper is as follows. Section 2 introduces the setting of *prediction problems*, in which the goal is to predict an unknown outcome $y \in Y$ from observables $x \in X$. A standard approach for learning a prediction rule $f : X \to Y$ is to specify a set of permissible mappings $\mathcal{F}$ and search for the mapping $f \in \mathcal{F}$ that best fits the available data. The key question is how to choose $\mathcal{F}$. Machine learning typically chooses large classes $\mathcal{F}$ that are flexible enough to approximate complex relationships between observables and outcomes, while remaining computationally tractable to optimize over. By contrast, economic modeling typically begins by restricting attention to a smaller $\mathcal{F}$ that embeds interpretable structure from the outset. As a result, machine learning algorithms approximate complex realities at the cost of interpretability, while economic modeling tends to preserve interpretability at the cost of failing to capture all the regularities in the data.

Sections 3–6 form the core of the paper and explore different ways to integrate flexible machine learning approaches with more structured economic modeling. Section

3 argues comparing the performance of economic models against machine learning algorithms can be illuminating about the properties of our existing models—in particular, how predictive, restrictive, and portable they are. This diagnostic step is important because we cannot improve our models without first understanding their limitations.

Section 4 goes further and shows how machine learning can help improve existing models or generate new ones. The methods discussed here seek to uncover regularities our current models miss—either by studying how a more predictive black box algorithm beats the economic model, or by computationally probing the contours of the economic model and discovering where its predictions diverge from reality. The ideal output is a new model that retains interpretability while achieving greater predictive accuracy.

Section 5 generalizes the notion of an economic model to include hybrid approaches that are part economic theory and part black box. In particular, machine learning can be used to flexibly recover model inputs from high-dimensional data, while respecting the constraints that the economic theory imposes for how those model inputs are mapped into outcomes.

Section 6 turns to large language models (LLMs). Early applications include using LLMs as substitutes for human subjects in experiments and as heuristic tools for exploring the space of ideas. I discuss how LLMs are qualitatively different from the supervised machine-learning approaches used in earlier papers, and point out some of the conceptual and methodological challenges they raise.

Section 7 discusses related work in machine learning, in particular the literature on interpretable machine learning and on whether LLMs have learned "world models." The motivation for this work is related to that of economic modeling, but there are also some important differences that I mention. Finally, Section 8 concludes with a discussion of open questions and directions for future research.

A number of excellent surveys examine the intersection of machine learning and economics, including Mullainathan and Spiess (2017), Athey (2019), and Athey and Imbens (2019) among others. These contributions have shaped the field's understanding of how machine learning can improve prediction, inference, and policy evaluation in economics. This article differs in its focus—rather than treating ML as a tool for empirical analysis, I consider how machine learning can be used to improve economic modeling.[1]

[1]This article is also distinct from a literature that takes an underlying causal model as given—

2 An Organizing Framework

Section 2.1 introduces a framework that encompasses common approaches in both machine learning and economic modeling. This framework is by no means exhaustive of questions in economics or of economic models,[2] but it will focus our attention on a subset of problems and models where comparing machine learning and economic modeling is particularly productive. Section 2.2 reviews the practice of out-of-sample testing; readers familiar with this topic can skip ahead.

2.1 Prediction Problems

An analyst seeks to predict an unknown outcome $y \in Y$ given an observable covariate vector $x \in X \subseteq \mathbb{R}^d$, where x and y are related by an unknown joint probability distribution P. This includes both observational settings and experimental settings, where in the latter the analyst controls the distribution of x.[3]

Example 1 (Predicting Certainty Equivalents for Lotteries). An analyst wants to predict how individuals value uncertain prospects. Specifically, for a lottery $\{(z_i, p_i)\}_{i=1}^n$ paying z_i with probability p_i, what certain payment y would an individual consider equivalent to the random outcome of the lottery? The features x describe the lottery's prizes and probabilities, and the outcome y is the *certainty equivalent.*

Example 2 (Predicting a Distribution of Play in Games). An analyst wants to predict how individuals will play in a given normal form game. A game is specified by a set of players $i = 1, \ldots, n$, a finite action set A_i for each player i, and a payoff matrix $u : \prod_{i=1}^n A_i \to \mathbb{R}^n$ mapping action profiles into payoff vectors. The outcome y is aggregate play in this game, i.e., a probability distribution over action profiles.

Example 3 (Predicting Information Diffusion). An analyst seeks to predict the extent of information diffusion in a social network following the initial seeding of informa-

---

formalized through potential outcomes, structural equations, or causal graphs—and develops machine learning methods to estimate causal parameters, heterogeneous treatment effects, or optimal policies (Athey and Imbens, 2016; Athey, 2019; Chernozhukov et al., 2017).

[2]Some economic models are of a fundamentally different category from machine learning; for example, some seek to distill relationships among concepts rather than making testable predictions about economic outcomes (Rubinstein, 2006). This paper will instead exclusively consider testable economic models, which take fit to data seriously as a goal and are extended or modified when researchers identify disagreements between their predictions and actual behavior.

[3]For example, if as in Example 1 the covariate vectors x describe lotteries and the outcome y is a certainty equivalent, then the experimenter chooses the lotteries that are presented in the experiment.

tion that a certain product is available among a subset of individuals. The feature vector x summarizes the structure of the underlying network and the identities or characteristics of the initially seeded individuals. The outcome y is the fraction of agents in the network who have heard about the product after T periods.

To select a *prediction rule* $f : X \to Y$, the analyst collects data $D = \{(x_i, y_i)\}_{i=1}^n$ and evaluates how well candidate functions f fit that data as measured by a loss function $\ell : Y \times Y \to \mathbb{R}$. The rule that minimizes average loss on the data is

$$f^* \in \arg\min_{f:X\to Y} \frac{1}{n} \sum_{(x,y)\in D} \ell(f(x), y) \tag{1}$$

where $\ell(f(x), y)$ quantifies the error to predicting $f(x)$ when the true outcome is y.[4]

An important consideration is what set of candidate rules f to search over in (1). Rather than optimizing over all possible maps $f : X \to Y$ as in (1), the analyst typically restricts attention to a specific class of functions $\mathcal{F}$ and solves:

$$f^* \in \arg\min_{f\in\mathcal{F}} \frac{1}{n} \sum_{(x,y)\in D} \ell(f(x), y) \tag{2}$$

This returns the prediction rule *within* $\mathcal{F}$ that maximizes fit to the data. Thus if $\mathcal{F}$ is restricted to prediction rules with a particular structure—say, linear or monotone in x_1—the prediction rule f^* will share that structure.

Traditional approaches in machine learning and economics differ in how the set $\mathcal{F}$ is chosen.

The machine learning choice of $\mathcal{F}$. In an idealized world with unlimited data and computing power, an analyst focused solely on accuracy would place no ex-ante restrictions on f, giving the data complete freedom to speak. In reality, both data and compute are limited. But modern machine learning's success lies in identifying function classes $\mathcal{F}$ that are not only highly expressive—i.e., capable of approximating complex relationships—but also amenable to efficient optimization. Algorithms search within these large classes to find the function that best fits the data.

Two example machine learning algorithms are neural networks and random forests. The architecture of a neural network—depicted in Figure 1—consists of successive layers of interconnected "neurons." Each neuron transforms the outputs from the

[4] A minimizer exists if the loss function is continuous and the relevant function space is compact: the space of all prediction rules in (1) or the class $\mathcal{F}$ in (2).

preceding layer by applying a nonlinear activation function to a weighted sum of those outputs. The network successively propagates these transformed outputs through multiple layers, and in this way is able to accommodate increasingly complex functions of the original inputs.

Neural networks are highly expressive: a network with even a single intermediate layer can approximate any continuous function on a compact domain, provided it has sufficiently many neurons (Hornik et al., 1989). In this respect, the representational capacity of neural networks is similar to that of nonparametric estimators.[5]

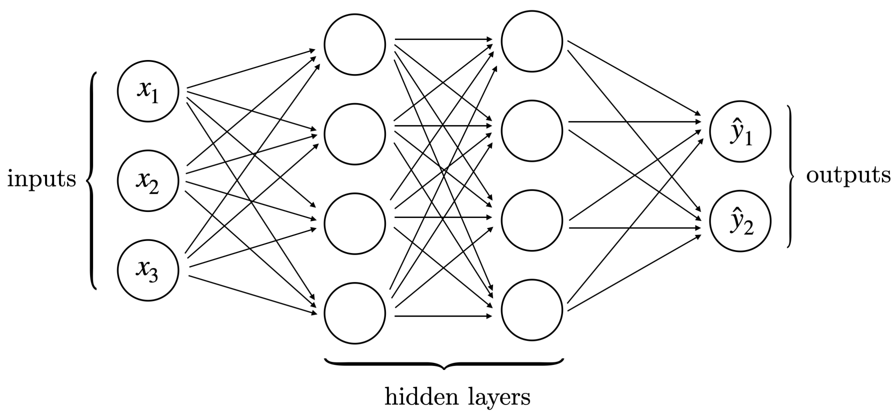

Figure 1: Example depiction of a neural network: the network processes input variables (leftmost layer) through a sequence of transformations, ultimately aggregating the final variables into predictions (rightmost layer).

The random forest algorithm—depicted in Figure 2—employs a different architecture but is also highly flexible. A random forest combines the predictions of an ensemble of decision trees, where each tree recursively partitions the feature space X, and assigns a constant prediction of y to each resulting region—typically the average outcome for regression or the most frequent class for classification. The ensemble is constructed by repeatedly drawing bootstrap samples from the data, and training a separate decision tree on each sample.[6] The random forest's prediction at a new input is then the average of the predictions across trees.

Both algorithms can be applied to the problem in Example 1 given data on lotteries and valuations. In this setting, a neural network would take the description of

[5]Deep networks achieve similar representational power more efficiently by stacking multiple layers.

[6]To further reduce correlation among trees, it is common to consider a random subset of features at each split.

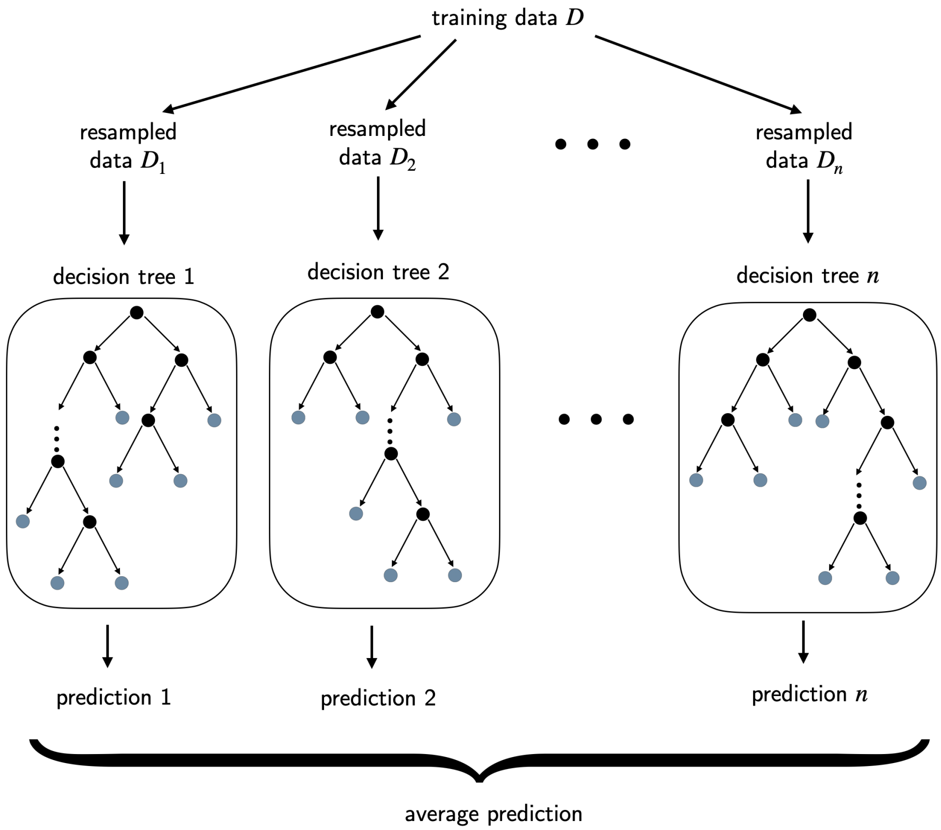

Figure 2: Depiction of a random forest algorithm.

each lottery—i.e., the vector of prizes and their probabilities—and learn a nonlinear mapping into intermediate features, which are then transformed through successive layers to produce a prediction of the certainty-equivalent. Ideally, these intermediate features would be interpretable (e.g., capturing the lottery's expected value or variance), but in practice they are typically not. A random forest, by contrast, would iteratively partition the space of lotteries into regions defined by simple rules (e.g., "the lottery does or does not have a negative prize") and assign to each region a predicted certainty equivalent. The final prediction averages across many such trees.

The use of these algorithms requires no prior knowledge about risk preferences or certainty equivalents, and yet with sufficient data they will produce highly accurate predictions. On the other hand, because the prediction rules learned by these algorithms are typically complex—multi-layered with many neurons per layer in the case of the neural net, and an ensemble of deep trees in the case of the random forest—they are difficult to interpret. While there are methods for deriving post-hoc explanations of specific predictions or aspects of a model's structure (see Section 7.1), the alternative approach in the following section begins by selecting a class of prediction rules that are individually interpretable by design.

The economic modeler's choice of $\mathcal{F}$. Sets $\mathcal{F}$ associated with economic models typically include only those prediction rules that possess some pre-specified structure, thus guaranteeing that the selected prediction rule is interpretable.

Consider the setting of Example 1. A standard model of risk preferences assumes that agents are expected utility maximizers with constant relative risk aversion (CRRA), henceforth EU–CRRA, defined over positive outcomes. Under this interpretation, the corresponding class $\mathcal{F}$ consists of prediction rules $f : \mathbb{R}^n_+ \times [0,1]^n \to \mathbb{R}$ of the form

$$f_\alpha(z_1, p_1, \ldots, z_n, p_n) = \left(\sum_{i=1}^{n} p_i z_i^\alpha \right)^{\frac{1}{\alpha}}, \tag{3}$$

where the parameter $\alpha \in \mathbb{R}_+$ indexes the curvature of utility and determines the agent's degree of risk aversion.[7] The parameter α is estimated from the data, and the selected prediction rule f_{α^*} corresponds to the best-fitting estimate α^*.

Relative to the machine learning algorithms discussed above, this class $\mathcal{F}$ imposes several restrictions on behavior. For example, EU-CRRA rules out the possibility that individuals misperceive probabilities, and that the sign of the prizes z_i affects the size of the risk aversion coefficient α. If actual behavior has those properties—so that EU-CRRA is misspecified—then this model will not fit real data perfectly. On the other hand, the selected prediction rule is guaranteed to have a clear and economically interpretable meaning: the agent maximizes their expected utility, where utility is a concave function of the dollar prize and α^* parametrizes the risk aversion of the agent.

These differences between EU-CRRA and the random forest illustrate a general distinction between economic modeling and machine learning. The later sections ask whether algorithms such as the neural network or random forest can help us to better understand and build upon economic models such as EU-CRRA.

[7] CRRA preferences are formally defined over total wealth; our specification either interprets z_i as final wealth or abstracts from background income.

2.2 Out-of-Sample Prediction

The machine learning algorithms discussed above are (intentionally) very flexible. This allows the algorithms to search over a wide range of possible relationships between inputs and outputs, but it also raises the risk of overfitting—producing a prediction rule that performs well on the data at hand but poorly on new, unseen data. To evaluate how well a model generalizes beyond the specific observations it was trained on, assessment of the performance of machine learning algorithms relies on *out-of-sample prediction*.

In the simplest setting, out-of-sample prediction begins by splitting the available data into two disjoint sets: a training set and a test set. The training set is used to fit, or "train," the model—that is, to select the specific prediction rule $f \in \mathcal{F}$ that solves (2). The selected prediction rule is then evaluated on the test set: for each observation (x_i, y_i) in the test set, the trained model produces a prediction $\hat{y}_i = f(x_i)$, and the prediction error $\ell(\hat{y}_i, y_i)$ is computed according to the chosen loss function. Because the test set was not used in fitting the model, this error can be used to estimate of the model's expected performance on new data drawn from the same distribution.

A common variation on the single train-test sample split is *K-fold cross-validation*. Here, the data is split into K roughly equal parts ("folds"). The model is trained on $K - 1$ folds and evaluated on the remaining fold, and this process is repeated K times so that each fold serves as the test set exactly once. The resulting K test errors are averaged to produce an overall estimate of out-of-sample performance. Cross-validation provides a more efficient use of limited data while still guarding against overfitting.[8]

From the perspective of economic modeling, out-of-sample prediction plays a role analogous to testing a theory on a new dataset rather than on the one used to estimate it. For flexible machine learning algorithms such as neural networks and random forests, the discipline imposed by out-of-sample testing is essential. This approach can also be useful for evaluating economic models, especially ones whose restrictiveness is not apparent (see the discussion in Section 3.2). The papers discussed in the following sections will use out-of-sample prediction to evaluate the predictive performance of economic models and machine learning algorithms alike.[9]

[8]See Bates et al. (2024), Fava (2025), and Lei (2025) for recent results on asymptotic theory for cross-validated errors.

[9]It is not strictly necessary to use out-of-sample testing to evaluate economic models with just a few parameters, and indeed the preceding literature has typically used in-sample estimators for the model's expected prediction error.

3 Comparing Economic Models and Machine Learning Algorithms

We expect economic models to be less predictive, more restrictive, and more portable than flexible machine learning algorithms. But is this actually the case? Assessing how well our models perform on these three dimensions is essential for understanding their strengths and limitations. This section introduces three computational metrics designed to evaluate these properties in economic models.

3.1 Predictive Accuracy

In most economic prediction problems, economic models fall well short of perfect accuracy. This can happen for two fundamentally different reasons. First, the model itself may fail to capture important relationships between the measured inputs and the outcome. For example, in the setting of Example 1, if individuals exhibit different risk preferences for losses and gains, or if they systematically misperceive probabilities, then EU–CRRA will not predict certainty equivalents exactly.

A very different possibility is that the measured inputs themselves have inherently limited predictive power. For instance, if presenting the same lottery in different ways leads individuals to report different certainty equivalents, then *no* model based solely on the lottery's characteristics could predict the reported value perfectly all the time.

These two sources of error call for different approaches to model improvement. If a model performs substantially worse than the best achievable accuracy for its given covariates, then one might look for new models based on the same inputs. For example, Kahneman and Tversky (1979)'s Cumulative Prospect Theory relaxed Expected Utility to allow for misperceptions of probabilities and different treatment of gains and losses, and achieved better predictions given the same inputs.

By contrast, if the model is already close to the best possible performance for the existing feature set, then it is futile to attempt to improve prediction by specifying new functional forms based on the same inputs; better predictions require identifying and measuring new variables. In the context of Example 1, this might mean letting the model take as input the subject's demographic covariates or the framing of the lottery.

One way to distinguish between these two error sources is to estimate the theoretical predictive limit given the measured covariates, i.e., the error of the prediction rule

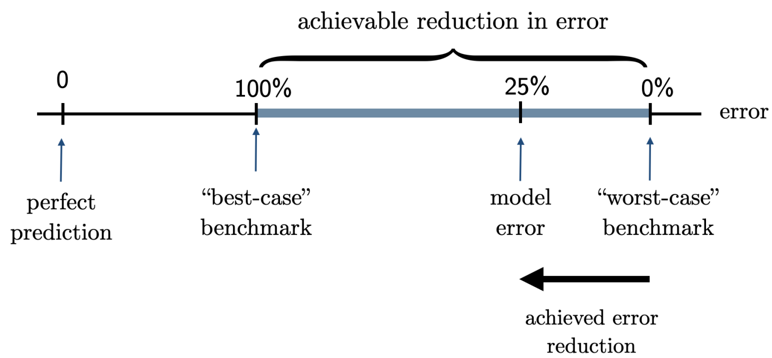

Figure 3: A model's *completeness* is the proportion of the gap between the worst-case and best-case prediction errors that the model is able to eliminate (Fudenberg et al., 2022). In the figure, the model's completeness is 0.25, indicating that the model reduces 25% of achievable reduction in error from the worst-case to the best-case benchmark.

that solves the unconstrained problem (1). This insight underlies Peysakhovich and Naecker (2017) and Fudenberg et al. (2022). Interestingly, for this goal, the flexibility of machine learning algorithms is not a drawback but rather an advantage. In fact, when the data contains sufficiently many (x, y) observations for each distinct value of x, then a simple "lookup table" algorithm—essentially a nonparametric estimator that learns the best prediction of y for each x separately—will approximate the best achievable error of any prediction rule $f : X \to Y$. In the limit of infinite data, such an algorithm achieves a completeness of 1. It thus provides a natural best-case benchmark for evaluating how much room remains for improvement over an economic model.

Fudenberg et al. (2022) further proposes using the error of a naive model (such as the expected value of the lottery in Example 1) as a worst-case benchmark, and then measuring where the economic model falls between these extremes. They define the *completeness* of a model as the fraction of the gap between the worst-case and best-case benchmark that the model closes,

$$\text{Completeness} = \frac{\text{Err}_{\text{baseline}} - \text{Err}_{\text{model}}}{\text{Err}_{\text{baseline}} - \text{Err}_{\text{best}}}$$

as illustrated in Figure 3. See Fudenberg et al. (2022) for an estimator of this theoretical quantity based on cross-validated estimates of individual errors, and Fudenberg et al. (2026) for the corresponding asymptotic theory.

Although lookup table algorithms can estimate the best achievable error for a diverse range of experimental data sets—where experimenters control which problems are presented to subjects and can elicit many observations of y for each x—they are not appropriate for most observational datasets. An alternative approach is to simply compare the performance of the economic model with that of another flexible machine learning algorithm, interpreting the latter as a proxy for the best achievable error in the problem. Several papers that do this find that economic models are nearly complete in their data. These settings include predicting certainty equivalents for two-outcome (Fudenberg, Kleinberg, Liang, and Mullainathan, 2022) and three-outcome lotteries (Peysakhovich and Naecker, 2017); predicting choice between risky outcomes with two (Ellis, Kariv, and Ozbay, 2023a) and three (Ellis, Kariv, and Ozbay, 2023b) equiprobable states of the world; and predicting the average cooperation rate in play of the repeated Prisoner's Dilemma (Fudenberg and Karreskog Rehbinder, 2024). These results do not rule out the possibility that machine learning algorithms would outperform the economic models given more data (although robustness checks such as varying the amount of data given to the machine learning algorithms can give us a sense of how valuable additional data would be). But they suggest that the economic models do not miss important regularities in these problems.

Other papers find that our existing models fall far short of the predictive accuracy of machine learning algorithms. The problems they consider include predicting the certainty equivalent for lotteries with unknown probabilities (Peysakhovich and Naecker, 2017); predicting the distribution of initial play of a normal-form game (Hartford, Wright, and Leyton-Brown, 2016); predicting repeated play of a normal-form game (Hirasawa, Kandori, and Matsushita, 2022); and predicting the distribution of choices between pairs of lotteries (Peterson, Bourgin, Agrawal, Reichman, and Griffiths, 2021), among others. It is an open question whether the problems and datasets for which economic models are complete can be systematically differentiated from the ones in which they are not. A natural conjecture is that economic models perform best in environments that are cognitively and strategically simple, where their abstractions are closer to literal descriptions of behavior. As environments grow more complex, behavior may increasingly reflect biases, misoptimization, and context dependence that flexible machine-learning methods can exploit, but that standard economic models typically abstract from.

It is important to note that these results reflect performance in specific prediction tasks and specific problem instances (e.g., specific lotteries or games), and a model's success in one domain does not necessarily translate to others. For example, a model that does well at predicting certainty equivalents for binary lotteries may perform less well with lotteries that have a larger support. Moreover, comparing economic models to machine learning algorithms in specific prediction problems presents an incomplete picture of performance, as economic models are intended to be useful across qualitatively different problems. We discuss this further in Section 3.3.

There is thus need for further research that compares economic models and machine learning algorithms beyond narrow prediction tasks. This will require new comparative metrics that more comprehensively evaluate economic models against black box algorithms, such as average performance or variation in performance across a broad range of prediction problems. It will also require empirical work that more systematically explores the problem space, revealing which types of problems our models handle well, and where they fall short. Section 4.1 discusses how algorithms themselves might help guide this exploration.

3.2 Restrictiveness

The completeness metric from the preceding subsection quantifies how well a model predicts real data relative to the best achievable performance given the model's inputs. Yet completeness alone cannot tell us whether this success is because the model precisely captures the regularities inherent in this particular data, or because the model is very flexible and can (with sufficient data) flexibly adapt to capture most patterns. Indeed, machine learning algorithms generally achieve high completeness for precisely the second reason. To tell these cases apart, we need a complementary measure that captures how restrictive a model is.

Selten (1991) proposed a measure for the restrictiveness of an economic model based on the fraction of datasets it can perfectly explain. In the language of Section 2, this is the fraction of all mappings $f : X \to Y$ that fall within the model's $\mathcal{F}$. A model that can explain any such mapping, given the right choice of parameter values, is completely unrestrictive.

While intuitive, there are two challenges with applying this measure. The first is that it is generally difficult to determine whether a particular f is consistent with a theory unless we already understand the theory's empirical content (e.g., by way of a representation theorem). Second, the Selten (1991) measure assigns identical scores to models that differ substantially in the kind of structure they impose. For a simple example, suppose there is a binary covariate $x \in \{x_0, x_1\}$ and an outcome

$y \in [0,1]$. Figure 4 compares two models: Model A allows for all mappings f that satisfy $f(x_1) > f(x_0)$, while Model B discretizes $[0,1]^2$ into a grid and allows for all mappings f where the pair $(f(x_0), f(x_1))$ falls into the shaded set. In Selten (1991)'s metric, both models are equally restrictive (allowing for half of the possible mappings), yet Model A captures a clear structural regularity while Model B does not.

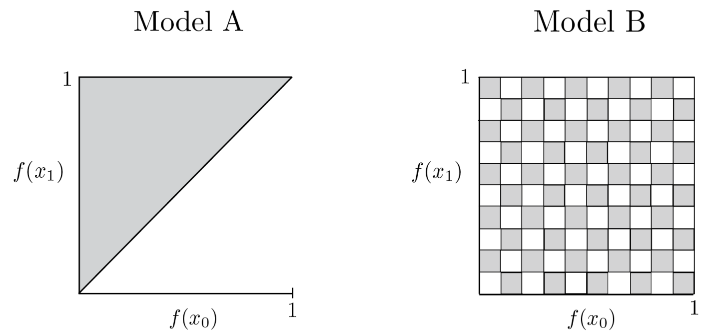

Figure 4: Models A and B are equally restrictive by the Selten (1991) measure.

Fudenberg et al. (2026) propose instead assessing restrictiveness as the average error a model makes when fitting randomly generated mappings f. They compute the model's best achievable fit to each of these mappings—formally, the smallest distance between the mapping and some $f \in \mathcal{F}$. These best-fit errors are averaged across many random draws and then normalized by the corresponding average error of a restrictive baseline model (such as predicting a constant). The resulting score lies between zero and one: a score of zero indicates that the model can perfectly approximate any mapping, while a score of one indicates that it approximates random mappings no better than the baseline.

Different from other classic measures of model restrictiveness such as the VC dimension,[10] this measure does not require an analytical characterization of the model's implications and can thus be more easily applied to assess a wide range of models. In particular, Fudenberg et al. (2026) apply their restrictiveness measure to evaluate a popular parametrization of Prospect Theory, which has been shown to have high completeness in predicting certainty equivalents for binary lotteries (Peysakhovich

[10]The VC dimension is defined for binary classification problems as the largest integer n such that there exists a set of n inputs that the model class can *shatter*, i.e., realize all 2^n possible binary labelings (Vapnik and Chervonenkis, 1971).

and Naecker, 2017; Fudenberg et al., 2022). The model's restrictiveness on the domain of binary lotteries turns out to be low (a normalized score of 0.28). This means that the model's good fit to certainty equivalent data for binary lotteries should not necessarily be taken to mean that the model identifies the right restrictive structure, as it could simply be due to the model's flexibility. On the other hand, as the size of the lottery's support expands the restrictiveness of the model increases, implying that a good fit to data from these more complex lotteries would provide stronger evidence that it captures the correct underlying structure.

Paired with completeness, restrictiveness locates models in a two-dimensional space, and helps identify a Pareto frontier of models that are undominated in both respects. This joint view makes it possible to distinguish between predictive success that reflects meaningful theoretical structure and predictive success that may simply reflect flexibility. In particular, since adding a free parameter to a model necessarily improves completeness and reduces restrictiveness, researchers can compare the loss in restrictiveness to the improvement in predictive accuracy to assess whether the latter compensates for the former, as in Ba et al. (2023).

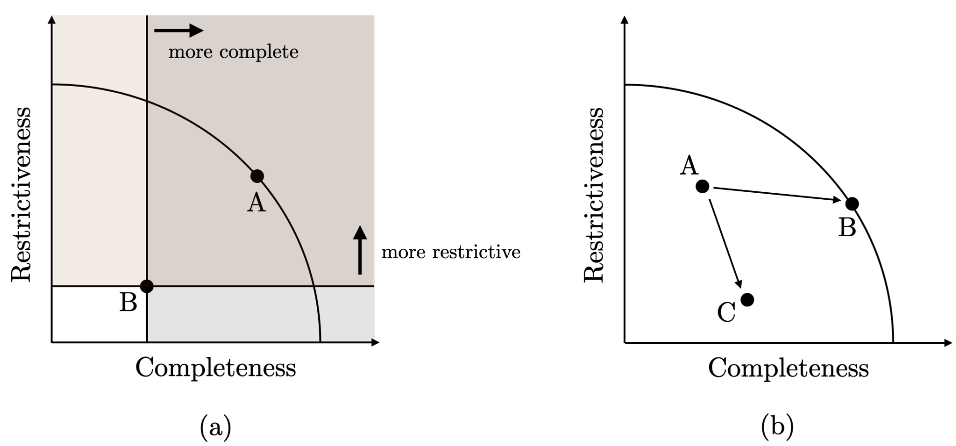

Figure 5: (a) Model A Pareto-dominates Model B by achieving both higher completeness and higher restrictiveness; that is, it simultaneously rules out more behaviors and also better explains real data. The curved line represents the Pareto frontier of models that are undominated in both dimensions. (b) Adding a free parameter to a model always involves a loss in restrictiveness and an improvement in completeness, but (all else equal) we prefer larger improvements in completeness and smaller reductions in restrictiveness, e.g., extending Model A to Model B rather than to Model C.

Figure 5 suggests an inherent tradeoff between restrictiveness and completeness.

But in the extreme case where the true mapping f for the prediction problem is known, there is no real trade-off to consider: the singleton class $\{f\}$ is at once maximally restrictive and fully complete. A tension between the two objectives instead emerges from other modeling challenges. First, the modeler may be uncertain about the data-generating process, and so must choose a larger hypothesis class $\mathcal{F}$ that has a good chance of containing the true f, even if this comes at the cost of greater flexibility. Second, the modeler may seek a model that applies across multiple domains, where the most predictive rule differs across domains (e.g., if different subject populations each behave as risk-averse expected utility maximizers, but with population-specific risk-aversion parameters). In this case the modeler may seek the minimal set $\mathcal{F}$ that includes the true f for each domain. The next section considers more explicitly the goal of finding a model that predicts well across economic environments.

3.3 Portability

To this point, we have evaluated models within a single environment. Yet a central goal of economic modeling is to uncover structure that connects distinct economic domains. Formally, recall from Section 2 that we have so far considered an unknown joint distribution for (X, Y) relating observables to an outcome of interest. One can think of *domains* as different settings in which that joint distribution varies. One might then ask: If a model was estimated on data drawn from one domain $(X_i, Y_i) \sim P$, how would it perform for predicting Y given X in a new domain where $(X_i, Y_i) \sim Q$?

A large machine learning literature addresses this challenge of *out-of-domain generalization* (see Zhou et al. (2021) for a survey). Unlike standard out-of-sample testing—where training and test sets are disjoint but drawn from the same distribution—out-of-domain generalization examines performance when training and test data come from different distributions. This is especially relevant for economics, where predictions are often needed in settings with no directly available data. For example, a development economist may need to forecast microfinance take-up in one Indian village using data from others, or an insurance economist may estimate willingness-to-pay for one set of insurance plans based on demand for a different set.

To systematically evaluate such cross-domain performance, Andrews et al. (2024) propose a general-purpose framework, which is applicable to both structural economic models and machine learning algorithms. They define a prediction method's *transfer error* to be how well it performs when trained in one (randomly drawn) economic domain and used to make predictions in another (randomly drawn) domain. The approach requires the analyst to collect a meta-dataset consisting of samples across multiple domains. By pairing training and test samples from the observed domains, the analyst can estimate expected performance in a new unobserved domain. The authors show how to construct simple confidence intervals for transfer performance under an assumption that the distributions governing the different domains are exchangeable, and further generalize these results when exchangeability is violated (but not by too much).

Applying this approach to models of risk preference, the authors find that black-box algorithms outperform economic models *within domain* but have worse transfer performance across domains. Moreover, the economic model does not uniformly improve on the black box algorithm, but is instead advantaged precisely when extrapolation is required, i.e., when the training and test sets involve different lotteries. This finding echoes Gechter et al. (2019)'s result that structural models produce better policy recommendations for conditional cash transfers in new contexts than black-box methods do. While Andrews et al. (2024) and Gechter et al. (2019) provide evidence that economic models capture generalizable structure, additional empirical work is needed to establish how broadly this result applies.

Portability thus joins completeness and restrictiveness as a third measurable axis. Completeness captures the share of explainable variation a model accounts for within a domain; restrictiveness measures how much it rules out; and portability quantifies how well these gains carry across domains.

4 Using ML to Generate New Models and Hypotheses

When machine learning algorithms outpredict our existing models, can we use these algorithms to uncover new regularities in behavior and to guide the extension of existing models? Several papers have made progress in exploring how algorithms can support the generation of new economic models and hypotheses. Below I discuss the successes and limitations of the ideas thus far.

4.1 Adversarially Breaking Existing Models

To improve on a model, we should know where it goes wrong.

A powerful approach from computer science is to adversarially train an algorithm to "break" an existing model. For example, Generative Adversarial Networks (GANs) are a class of machine learning models that pit two neural networks—a generator and a discriminator—against each other in a game. The generator's task is to create synthetic data (such as images) that mimic real data, while the discriminator's role is to distinguish between real and generated data. As the generator improves at creating realistic data, the discriminator gets better at identifying fakes, driving both algorithms to improve over time.

This adversarial approach is used in Fudenberg and Liang (2019) to guide experimental design. They illustrate their method on Stahl and Wilson (1994)'s Level-1 model of play in normal-form games. Although this model predicts existing experimental data very well, its good performance on these games does not imply that its performance will generalize to other games, since the games selected for lab experiments may share some special structure. To address this, Fudenberg and Liang (2019) use machine learning to identify games on which the Level-1 model may fail. Using the existing experimental data, they train an algorithm that takes a game matrix as input and predicts the Level-1 model's predictive accuracy on that game. They then use the trained algorithm to identify a large set of new games for which the Level-1 model is predicted to have low predictive accuracy, and collect experimental data on behavior in these games. They find that the Level-1 model indeed predicts behavior poorly in these games, and show that behavior in these games is not random but instead follows consistent patterns better explained by an alternative model.

While Fudenberg and Liang (2019) use machine learning to adversarially select cases where a theory is likely to break down, Mullainathan and Rambachan (2024) propose a complementary framework that treats model failure itself as the object of study. They frame such examples as *anomalies*—instances, like the Allais paradox, that serve as an exemplar of a kind of behavior that could not be explained by the theory, and thus illuminate the predictive boundary of the theory. They formalize the identification of anomalies as an adversarial game between a theory and a falsifier, where the falsifier proposes examples—i.e., collections of feature-outcome pairs—and the theory attempts to explain these examples by fitting its allowable functions to them. To solve for equilibrium of this game, they develop a gradient descent procedure that takes small steps in directions where the model's predictions change minimally, and yet actual behavior changes substantially. Applying this method to expected utility theory, the authors uncover known anomalies—such as the Allais paradox and the Common Ratio effect—as well as anomalies that have not been previously documented. They further verify that experimental subjects indeed violate expected utility theory on these algorithmically generated examples.

These papers demonstrate that machine learning can systematically expose the limitations of economic models, helping researchers refine theories by focusing on parts of the problem space where those models fall short. Designing environments that reveal a model's weaknesses can surface behavioral patterns that remain poorly understood. The challenge now is to extend these methods to moreover characterize the anomalies they uncover, enabling researchers to map not only where models break down, but also what they miss.

4.2 Combining Human Insight with Machine Learning to Uncover New Regularities

One way to understand what the machine learning algorithms have uncovered is to use human insight to interpret its edge over existing models.

One kind of human insight is researchers' own domain knowledge. For example, Fudenberg and Liang (2019) find that a machine learning algorithm outperforms the leading economic models in out-of-sample prediction. To understand why, they examine the subset of games where the machine learning algorithm is correct and the economic model's prediction is not. These games are interesting because: (1) the good performance of the machine learning algorithms suggests there is an underlying regularity, and (2) the poor performance of the economic models suggests that this regularity is not yet reflected in the model.

Examining these games reveals a common pattern: in each game, subjects responded to payoffs as if they were risk averse. The Level-1 model, which predicts that subjects choose the action that is a best response to uniform play by their opponent, does not allow for risk aversion, but can easily be extended to accommodate it. This new model matches the performance of the black box algorithms on the existing data, and continues to perform well on new data of play in other games, including in the later analysis of Külpmann and Kuzmics (2022).

Hirasawa et al. (2022) consider repeated play of a normal-form game with a unique (and complex) mixed strategy equilibrium, and use machine learning to reveal new behavioral regularities. They first compare the performance of Camerer and Hua Ho (1999)'s Experience-Weighted Attraction model (henceforth EWA) against that of machine learning models, finding that EWA is only approximately 30% complete. To understand which covariates play an important role in the machine learning algorithms, they employ two complementary approaches: examining feature importance measures (see Section 7.1) and analyzing which covariates are consistently selected by LASSO regularization across all rounds of cross-validation. Both approaches point to the same insight: players categorize the actions into two distinct classes and negatively autocorrelate their choices within each class—that is, if a player recently chose an action from one class, she is less likely to choose from that class again. The authors infer that if a player's own behavior exhibits this serial correlation, she may also recognize that her opponent is subject to the same tendency. Incorporating these insights into EWA, Hirasawa et al. (2022) derive an interpretable behavioral model whose predictive power is nearly comparable to that of the best-performing machine learning model.

Finally, Ludwig and Mullainathan (2024) address the challenge of generating new hypotheses, leveraging crowd wisdom to understand what the machine learning algorithm understands that the economic model does not. Specifically, they consider the problem of how judges make pre-trial detention decisions based on a defendant's mugshot. To better interpret the algorithm's predictions, they employ a morphing algorithm, which creates synthetic mugshots that transition smoothly between faces with low and high predicted likelihoods of detention. This morphing process systematically alters only those facial features that the algorithm deems relevant to its predictions, while holding other attributes constant. By generating pairs of synthetic images that differ mainly in their predicted detention probabilities, the authors are able to isolate facial characteristics that influence the algorithm's decisions. Human subjects are then shown these morphed images and asked to describe the differences they perceive, thus assigning interpretable labels to the algorithm's findings. This approach can be iterated further by generating pairs of images that are not distinguished by the first feature, and yet have different predicted likelihoods of detention. Through this process the authors discover that "well-groomed" and "heavy-faced" defendants are more likely to be released.

These approaches all yield interpretable new directions in which existing models can be improved, and thus provide a proof of concept that machine learning can guide discovery of simple and portable models. On the other hand, these papers do not fully automate the process of that discovery—human expertise and intuition remain an essential part of how the algorithm's output is transformed into meaningful hypotheses and models that can be tested further. An open question is to what extent the final step can also be guided by algorithms. The recent development of large language models (LLMs) offers an opportunity to close that gap. Are LLMs capable of translating the complex patterns and relationships discovered by black-box algorithms into human-interpretable explanations and testable models? This is an empirical question: Since LLMs are trained on the existing literature, their ability to generate explanations may be (for now) confined by the boundaries of existing knowledge. Nevertheless, the ability of LLMs to communicate in ordinary language gives them a distinctive and potentially significant role in the further automation of scientific modeling. (See Section 6 for further discussion.)

5 Hybrid Econ-ML Models

The preceding sections considered how machine learning can evaluate, critique, and extend traditional economic models. This section explores a different approach: rather than simply using machine learning algorithms to help us in the construction of economic models, we might broaden our conception of what a model is and create new "hybrid" models that combine economic structure with flexible, data-driven estimation.

There are multiple ways to implement such hybrid models. Section 5.1 focuses on settings in which machine learning is used to flexibly learn certain inputs to the model. These inputs—such as the "strength of a network tie"—are not directly observable, but can be predicted given other observables, such as how often two households interact or borrow money from one another. In these cases, machine learning methods are used to learn the mapping from basic observables to model inputs.

Section 5.2 reverses this logic. Rather than using machine learning to supply inputs to a fixed model, these approaches use economic theory to constrain the space of prediction rules explored by an algorithm. Such constraints may enter indirectly, by guiding feature construction, or more directly, by imposing structural restrictions on the algorithm's architecture.

5.1 Using ML to Flexibly Learn Model Inputs

We have so far considered models that directly relate a set of observables X to a set of outcomes Y. Many theories, however, involve inputs that are not directly observable, such as the utility of a dollar prize or a network of relationships across individuals. These theories specify how the inputs relate to the outcome, but leave open how the inputs are themselves derived from observables. This separation is often intentional, distinguishing the causal pathways we want to understand from the inputs we are willing to take as given. See Figure 6 for an illustration.

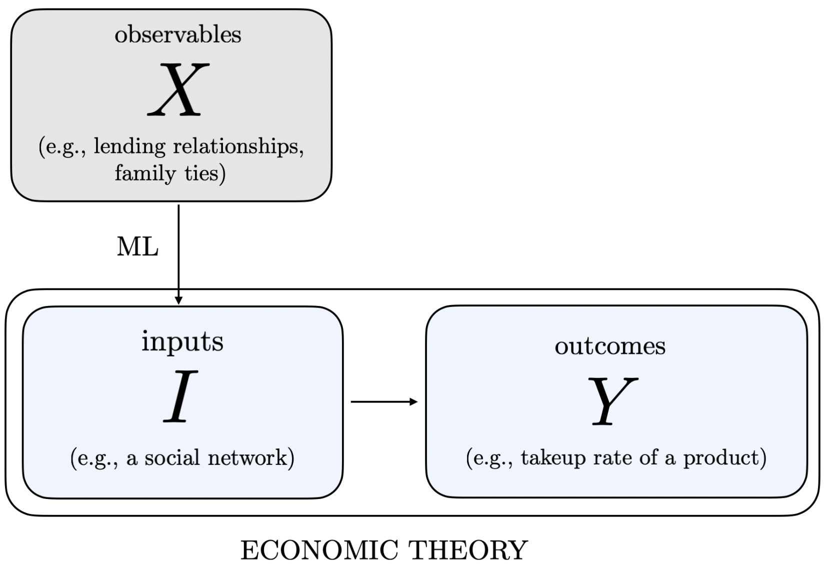

Figure 6: Some economic theories can be viewed as a set of inputs I and a set of maps $m : I \to Y$ relating inputs to outcomes. For example, a model of information diffusion might start with a social network as an input, and model how information is passed around the network, ultimately leading to a prediction of takeup rates. Such a model does not specify how we measure the social network, so to bring the model to data the researcher needs to further specify a map $\tau : X \to I$ taking measurable observables into model inputs.

To apply such a model to data, researchers typically specify a functional form that links observables to the model inputs. But the predictive performance of the model, as assessed in this way, depends both on the quality of the model and the quality of the measurement of the model inputs. If the model predicts poorly, one possibility is mismeasurement of the inputs, rather than a failure of the model. The following papers apply machine learning to flexibly learn these inputs from the data.

Risky choice. Peterson et al. (2021) study individual choice between lotteries. Two canonical economic frameworks for this problem are Expected Utility Theory (EU) and Prospect Theory (PT). Under EU, an individual evaluates a gamble that pays z_i with probability p_i according to $\sum_i u(z_i)p_i$ for some utility function u. Prospect Theory generalizes this by transforming probabilities through a weighting function π, evaluating lotteries as $\sum_i u(z_i)\pi(p_i)$.

In empirical applications of these models, it is standard to impose parametric forms for u and π, for example $u(z_i) = z_i^\alpha$ (as in Section 2) and $\pi(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}$ (as proposed in Goldstein and Einhorn (1987)). Peterson et al. (2021) instead allow a neural network to flexibly learn u and π from the data, referring to these specifications as neural EU and neural PT. Comparing neural EU with neural PT enables the authors to isolate the predictive gain from allowing probability weighting, without committing to any specific functional form for how probabilities are misperceived. They find that neural PT achieves slightly higher out-of-sample accuracy than neural EU, and also outperforms PT under parametric specifications of π that appear in the literature. A complementary direction for analysis would be to assess the restrictiveness of these models (Section 3.2), thereby quantifying not only the gains in predictive accuracy but also the loss in model restrictiveness.

Salience. Since Schelling (1960), the importance of salience in coordination games has been widely acknowledged, and yet rarely invoked in game theory due to the difficulty of measuring it. Li and Camerer (2022) show that machine learning can quantify the salience of different actions, and that this quantity can be input directly into game-theoretic models.

Consider hide and seek games in which two players—a hider and a seeker—are simultaneously asked to select a location from an image. The seeker wins if their locations match, and the hider wins otherwise. There is a unique Nash equilibrium in this game in which all locations are chosen equally often. But in empirical play, the seeker has an advantage relative to the equilibrium prediction, with matches happening more often than what uniform play would imply.

To explain this, Li and Camerer (2022) propose a variation on a cognitive hierarchy model (Nagel, 1995; Camerer et al., 2004; Crawford and Iriberri, 2007), which incorporates the idea of focal points. In this framework, level-k players believe opponents are drawn from lower cognitive levels according to a Poisson distribution and logit best-respond to these beliefs. The model is anchored by level-0 players, who are traditionally assumed to choose uniformly at random from the available actions. Li and Camerer (2022) instead suppose that level-0 players choose locations with probability proportional to their visual salience, as quantified by the Salience Attention Model—a deep neural network developed in computer vision that predicts which regions of an image will attract human attention (Vig et al., 2014; Cornia et al., 2018). This salience-augmented model better matches the empirical data.

Discrete choice. In discrete choice models, the accuracy of predicted consumer behavior depends critically on how the products are characterized, but it is not obvious what the relevant characteristics are.

Compiani et al. (2025) use pre-trained machine learning models to extract *embeddings*—i.e., high-dimensional vector representations—from unstructured product image and text data. They apply principal components analysis to these high-dimensional embeddings to recover the main dimensions of product differentiation, and incorporate these principal components in a mixed logit demand model (McFadden and Train, 2000; Berry et al., 1995). This richer representation leads to more accurate recovery of substitution patterns across products, which in turn improves out-of-sample counterfactual predictions, such as predicting which products consumers would choose from a new choice set.

Han and Lee (2025) adopt a related approach to study font markets and copyright policy. They represent each font as a point in a high-dimensional "visual space" derived from embeddings of font images, where distances in this space reflect perceived visual similarity. Firm competition is modeled as spatial differentiation within this space, with products that are closer together interpreted as closer substitutes. Within this framework, they model copyright protection as restricting entry to be at least a distance d away from any existing font, and characterize how increasing d affects firm entry decisions and consumer welfare.

Parameter heterogeneity. Farrell et al. (2021) considers settings in which an outcome y depends on treatment variables T through a model with parameters θ. They assume that the structural relationship between y and T is common across the population, but that the parameters θ may vary with individual characteristics X. For example, the outcome may be linear in T for all individuals, i.e., $y = \langle \theta, T \rangle$, while the coefficients θ differ across individuals. This generalization turns θ from a fixed vector into a potentially high-dimensional, nonlinear mapping $\theta(X)$, capturing heterogeneity not specified by the original theory.

To estimate these heterogeneous parameter functions, they design a *structural deep neural network* in which the hidden layers feed into a "parameter layer" that outputs $\theta(X)$. This parameter layer is then passed through the fixed structural form—what they call the "model layer." This architecture preserves the interpretability of the structural model, while exploiting deep learning's flexibility to recover rich, nonlinear patterns in parameter heterogeneity.

There are several advantages to these "machine-learning-augmented" theories. First, they let us bring a broader range of theories to data by operationalizing concepts—such as salience or visual similarity—that were previously too vague or complex to quantify, thus expanding the set of theories we can formally test.

Second, they make existing theories more predictive by leveraging sources of complex and unstructured data that were previously ignored. Using this data, we can measure our model inputs more accurately, thereby improving the theory's explanatory and predictive power.

Finally, these approaches help identify where further scientific progress is most needed. If optimizing the construction of inputs substantially improves predictive performance, this suggests that the bottleneck is measurement. And if, instead, increasingly sophisticated reconstruction of model inputs from data yields little predictive improvement, then it may be that the theory itself needs to be revised.

5.2 Using Economic Theory to Guide or Constrain ML

Reversing the perspective of the previous section, we now consider how economic theory can inform the design of machine learning algorithms.

Feature Engineering. One way theory can contribute is by informing the selection and construction of features for the algorithm, a process often referred to as feature engineering. Consider the problem of predicting certainty equivalents for lotteries (Example 1). Given sufficient data, a highly flexible algorithm should (in principle) discover any relevant summary statistics of the lottery for this problem, such as its expected value or variance. In practice, when the search space is large and data are limited, performance can be improved by using theory to identify meaningful combinations of features and adding them explicitly to the feature set (see Figure 7).

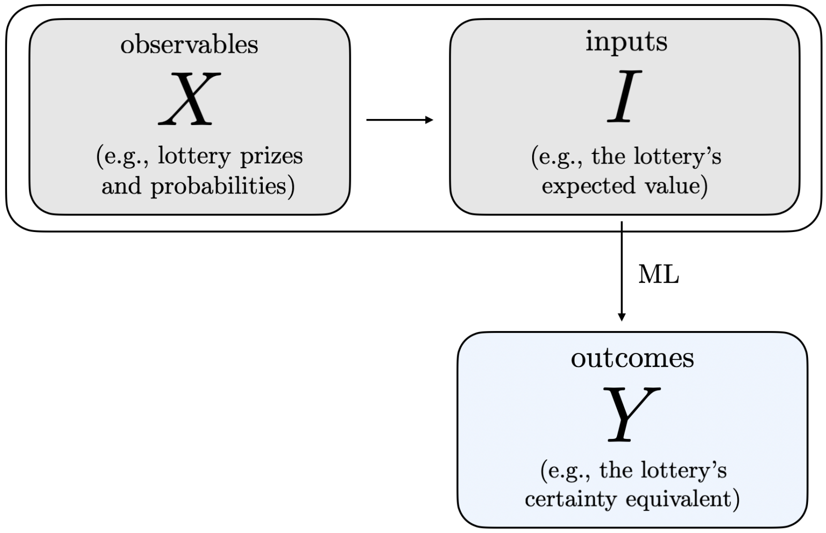

Figure 7: Some theories identify relevant concepts, without specifying a precise form in which those concepts predict the outcome.

Plonsky et al. (2017) do precisely this. The problem they consider is predicting choice between two lotteries. They compare algorithms trained on three nested feature sets: The first simply describes the two lotteries that are presented to the decision-maker, i.e., the prizes and their probabilities. The second feature set additionally includes summary statistics such as the difference between the lotteries' expected values, the difference between their standard deviations, and the difference between their minimal and maximal outcomes. The final feature set adds additional features based on the preceding theoretical literature. These include, among others, whether one lottery first-order stochastically dominates another, whether only losses are possible, and measures of regret.

The second and third feature sets do not introduce new information relative to the first: all of the features can be constructed from the lotteries' prizes and probabilities. Yet Plonsky et al. (2017) find that adding the summary statistics reduces mean squared error (MSE) by over 70%, and that adding the theory-motivated features yields a further 37% reduction. Thus in this case, there is substantial predictive value to using theory to identify relevant features.[11]

[11] A similar approach is used in Fudenberg and Liang (2019) and Wright and Leyton-Brown (2014) where domain knowledge is used to hand-craft features for predicting strategic play in matrix games.

Imposing Economic Theory as a Hard Constraint. Hartford et al. (2016) draw on economic theory to impose structural constraints on behavior and encode these constraints directly into a neural-network architecture.[12] In their application—predicting play in normal-form games—the network hard-codes symmetries implied by theory. For example, relabeling either player's actions leaves predicted behavior unchanged, apart from a corresponding permutation of probabilities. The network also implements iterative strategic reasoning through a sequence of action–response layers, with each layer updating choices based on beliefs formed in the previous one. The resulting model outperforms standard deep-learning baselines that rely on hand-crafted features (i.e., input variables manually constructed by the researcher).

Imposing Economic Theory as a Soft Constraint. Fessler and Kasy (2019) instead treat economic theory as a "soft constraint" by anchoring estimation to the predictions of a structural model. They study the problem of learning a parameter vector β. Two extremes are to rely entirely on an unrestricted estimator or to fully adopt the predictions of a structural model. Rather than choosing between these, they propose a weighted average, with weights determined by how well the model's predictions fit the data. Thus, instead of rigidly imposing the theory's predictions, this approach uses the predictions of economic models as an anchor for parameter estimation, shrinking estimators towards theoretical predictions when they are approximately correct, but reverting to an unrestricted model if the data diverge too much from the theory.

6 Large language models

Large language models (LLMs) represent a qualitative break from the machine learning tools discussed so far. First, in contrast to traditional supervised learning algorithms that are designed to map a specific input x to a specific output y (such as diagnosing a medical condition from an image scan), LLMs are "general purpose technologies" that are capable of responding coherently to most natural-language prompts.

A second important difference is that LLMs operate in natural language, and thus

[12] It is well established that hard-coding domain knowledge of the input, particularly its topological structure, can improve representation learning (Bengio et al., 2013). For example, Clark and Storkey (2015) exploit rotational invariance when training a Go-playing network.

can participate directly in the conceptual dialogue that drives theory construction. LLMs can process verbal arguments about causality, translate between mathematical formalism and economic intuition, and link abstract models to real-world phenomena. Thus, they have the potential to move beyond computing optimal solutions or estimating parameters, toward proposing mechanisms, identifying assumptions, and articulating the economic logic that connects primitives to predictions.

How best to use LLMs to augment economic modeling is an emerging area of research, with many of the most intriguing possibilities still being explored. Nevertheless, some unifying ideas are beginning to take shape. Section 6.1 presents LLMs as synthetic behavioral subjects capable of simulating human decision-making, and Section 6.2 examines their potential to automate certain heuristic and creative aspects of research. Together, these perspectives suggest promising ways to accelerate and improve economics research, while also introducing distinct methodological and conceptual challenges.

6.1 LLMs as Behavioral Subjects

Horton (2023) argues that LLMs can be viewed as "computational models of humans" that approximate human reasoning and biases sufficiently well to stand in as synthetic subjects in economic experiments.

This is appealing for several reasons. First, there is an obvious cost advantage relative to recruiting and paying human subjects. Second, the researcher can run a broader range of experiments, including (for now) ones that would be infeasible or unethical to conduct with human participants. For example, Aher et al. (2023) use LLMs to simulate an obedience study based on the classic Milgram shock experiment (Milgram, 1963).[13] Third, LLMs potentially offer unprecedented control over the subject pool: Researchers can systematically vary traits such as age, gender, cultural background, political affiliation, cognitive style, or even risk tolerance. While further research is needed to assess how successfully identity can be elicited and how precisely users can control this process, this approach has the potential to generate precisely tailored subject profiles, enabling researchers to query virtually any subject population on any question.

A critical question is whether LLMs in fact faithfully replicate human personality and decision-making. Several papers have conducted "Turing tests," asking whether

[13]The Milgram experiment asked participants to administer painful electric shocks to another person and found that many complied despite believing the shocks to be harmful.

LLM behaviors and human behaviors can be distinguished from one another. These papers find that LLMs broadly reproduce human tendencies.

For example, Aher et al. (2023) ask LLMs to participate in the ultimatum game, in which one player proposes how to split a $10 endowment, and the other decides whether to accept the proposal (where rejection leads to both players receiving nothing). The LLM distribution of offers and acceptance rates resembles human data, where in both cases offers of 50–100% are almost always accepted, while offers below 10% are almost never accepted.

Mei et al. (2024) extend this analysis to a wider set of experimental games, including (among others) the dictator game, trust game, and finitely repeated Prisoner's Dilemma. There is a large quantity of human data for these games,[14] and Mei et al. (2024) use these to assess the likelihood of the LLM responses under the empirical human response distribution. Specifically, they draw a random action from the LLM's response distribution and compare it to an action drawn at random from the empirical human distribution. They then ask: which of these two actions is more probable under the human distribution? Averaging across games, they find that the LLM-generated action is typically assigned higher probability, and is in this sense more "human-like." (This finding should be interpreted with care: a deterministic distribution on the modal human action maximizes this "human-like" metric, but genuine mimicry may also require capturing heterogeneity across the distribution; we return to this point below.) Moreover, like humans, LLMs are sensitive to context and framing: Mei et al. (2024) show that LLMs become more generous in the ultimatum and dictator games when told their choices will be observed by a third party, and Horton (2023) find that LLMs are more likely to choose an option when it is presented as the status quo.

Nonetheless, some systematic differences remain. In a wisdom-of-the-crowds experiment, Aher et al. (2023) instruct LLMs to predict the distribution of human responses to ten general-knowledge questions including "How many bones does an adult human have?". They find that the LLMs skew toward accurate answers, in one case predicting that a majority of simulated participants would answer all ten of the questions correctly. And across games with a distributional element (such as the dictator and ultimatum game), Mei et al. (2024) find that LLMs display greater cooperation and altruism than the median human. Both tendencies likely reflect post-

[14] Raman et al. (2025) evaluate LLMs on a battery of microeconomic questions, but their focus is on assessing its rationality rather than its similarity to human behavior.

training objectives set by the companies that develop these models; that is, after the initial training, developers often further adjust the model to prioritize being helpful and honest (see, for example, Bai et al. (2022)). These results therefore highlight that post-training can introduce predictable biases that shift model outputs away from the true distribution of human responses.

This line of research raises several fundamental questions.

What is the LLM's identity, and can we control it? A first challenge in using LLMs as experimental subjects is determining what identity—or identities—an LLM represents. At least three possibilities emerge. First, we might view the LLM as a single, representative agent, with its answers corresponding to the modal response in the human population. This view is consistent with Mei et al. (2024)'s description of heterogeneity across LLM responses as "within subject" variation, and with their evaluation of human-likeness based on how far LLM responses differ from a typical human. If so, deploying LLMs as experimental subjects could reveal aggregate behavior and patterns, but would offer limited insight into heterogeneity across people.

Second, the LLM could be viewed as representing the full human population, with each response a random draw from the distribution of human behaviors and preferences. While LLMs do exhibit some randomness in their responses (which is tunable via the temperature parameter), Mei et al. (2024) find that the variation across LLM responses falls far short of the heterogeneity observed among human subjects. The distributions also differ qualitatively: human response distributions have more distinct peaks, indicating multiple behavioral subgroups, whereas LLM response distributions are typically unimodal.

Third, we might view the LLM as a collection of identities, capable of mimicking a wide range of individuals when appropriately primed. Aher et al. (2023) and Horton (2023) explicitly assume this perspective and induce variation in LLM responses by priming different identities—Aher et al. (2023) by randomizing the names of the people whose responses the LLM is asked to simulate, and Horton (2023) by priming the LLM with different political identities. Mei et al. (2024), too, note the effects of identity priming: they find that in the ultimatum game, LLMs shift toward game-theoretically optimal strategies (accepting offers as low as $1) when asked to act as a mathematician, and demand $50 (out of $100) in most cases when prompted as a legislator.

The variation in how LLMs are conceptualized across these papers points to the need for a clearer account of the identities LLMs can meaningfully represent before they are used as experimental subjects at scale. One key dimension concerns the treatment of *heterogeneity*: should LLMs be used as representative agents, or can they reproduce the full distribution of human behavior? A second, related, dimension concerns *population structure*: can LLMs mimic any given identity, or are there limits? For example, we might expect that LLM performance degrades when asked to mimic highly idiosyncratic individuals or groups that are weakly represented in the training data.

Understanding the capabilities and limits of LLM imitation requires systematic methods for eliciting and validating LLM identities. Specifically, we need experimental designs that allow researchers to test hypotheses about how different prompting or conditioning strategies shape model behavior.

Can we trust LLM generalization? Suppose we establish in some principled way that LLM responses approximate human responses across a broad range of settings and data. A crucial question remains: will the LLM behave as humans do in new experiments, when confronted with problems and scenarios that are not represented in its training data?

This is critical to the value of LLMs for advancing behavioral science, as Horton (2023) envisions, since they are only useful if they can generate reliable responses for problems where we do not already have data. In fact, Horton (2023) proposes that LLMs serve as sandbox laboratories in which one can test the consequences of counterfactual worlds and conditions that would be difficult to credibly create with human subjects. He draws an analogy with economic theory, citing Lucas (1980):

> One of the functions of theoretical economics is to provide fully articulated, artificial economic systems that can serve as laboratories in which policies that would be prohibitively expensive to experiment with in actual economies can be tested at much lower cost.

But there is a crucial difference: the predictions of a well-specified theory are disciplined by the internal logic of the model. By contrast, LLMs offer no guarantee of such internal consistency. Their behavior in new tasks need not be (substantially) constrained by their behavior in prior tasks. In this respect, they resemble traditional machine learning models, and as with other ML systems, there is no reason to assume their predictions generalize.

It remains possible that an LLM has distilled some underlying representation of "what it is to be human," and that this representation enforces cross-context consistency. If so, the model's behavior in new environments could indeed remain human-like. But until we have evidence that its internal structure imposes such discipline, we should be cautious in interpreting out-of-sample LLM behaviors as indicative of what people will do. (Section 7.2 discusses related work examining whether LLMs have learned "world models"—structured representations of underlying rules—or are fundamentally limited to pattern-matching.)

Addressing LLM generalization requires progress on two fronts: developing methodologies for measuring LLM performance, and conducting empirical assessments across a broad range of applications. On the methodological front, Gao et al. (2026) propose a measure for the domain-specific value of LLMs as the amount of human data they substitute for. Specifically, they ask how much human data would be required for a conventional model (such as an economic model or a machine learning algorithm) trained on that data to match the predictive performance of the pretrained LLM in that domain, and define the LLM's *equivalent sample size* to be the smallest size of training data required for the model to match the LLM. This measure provides a principled way to quantify LLM usefulness that accounts for domain-specific variation.

On the empirical front, Anthis et al. (2025) conduct a systematic review of studies that use LLMs to simulate human behavior, finding that while LLMs often replicate directional effects observed in human studies, they frequently fail to match effect sizes and can exhibit systematic biases not present in human populations. One of these studies is Hewitt et al. (2024), which compiles an archive of 70 pre-registered, nationally representative survey experiments and prompts GPT-4 to simulate how representative samples of Americans would respond to the experimental stimuli. Predictions derived from simulated responses correlate strikingly well with actual treatment effects, equaling or surpassing the predictive accuracy of human forecasters. Notably, accuracy remained high for unpublished studies that could not have appeared in the model's training data, suggesting genuine predictive capability rather than memorization. This type of held-out assessment—where the LLM is asked to predict outcomes for experiments not represented in its training data—is a good strategy for testing whether models have learned generalizable principles of human behavior. See Ludwig et al. (2025) for further discussion of this point and guidance on using LLMs in empirical research.

6.2 LLMs for Creative and Heuristic Research Tasks

A different potential use of LLMs is to automate aspects of the research process that are informal and creative. This possibility may seem both surprising and provocative. Creativity has traditionally been regarded as a uniquely human capability, drawing a sharp boundary between human generation and automated work. Yet LLMs have demonstrated a capacity to generate outputs that resemble creative, associative thinking. Whether its capabilities in this domain are sufficient to make a substantive contribution to research is an empirical question.

Han (2025) highlights one possible application: identifying instrumental variables (IVs). An IV's validity is rarely established through formal proof but rather through persuasive argumentation about its relevance and exogeneity. LLMs can potentially automate this search. Han (2025) demonstrates in classic economic problems that LLMs can not only reproduce known instrumental variables (IVs) but also propose new, plausible candidates. Consider the problem of estimating the causal effect of years of schooling on earnings. A valid IV in this context should be predictive of schooling attainment (relevance) but unrelated to any unobserved determinants of earnings (exogeneity), thus affecting earnings only through years of schooling. When appropriately prompted, the LLMs recover standard instruments for this problem, including the distance from a student's home to the nearest college and the number of siblings simultaneously attending college. The LLMs also suggest novel candidates, including campus crime rates and campus housing capacity. In a poll, experts found these novel IVs to be somewhat less valid than the established ones, but not by a wide margin.

In assessing findings of this kind, a natural concern is whether LLMs "discover" classic IVs through genuine reasoning or simply recall them from the training data. The novel candidates it produces offer some evidence of creative capability, but the true test going forward will be whether it can generate high-quality IVs for genuinely new problems.

Another creative endeavor in research is the generation of novel hypotheses. We already considered this goal in Section 4, where we observed that machine learning algorithms could be helpful in directing human attention, but ultimately required human input into the final formulation and naming of the hypothesis. LLMs offer an opportunity to automate the process more completely.

Movva et al. (2024) propose a method for generating natural-language hypotheses using *sparse autoencoders*. An autoencoder is a neural network that learns to compress an input vector x into a lower-dimensional hidden vector h (the encoder) and then reconstruct it back into $\hat{x}$ (the decoder). It is trained so that $\hat{x}$ is as close as possible to x. This forces the model to discover structure in the data—features that are sufficient to reconstruct the original input. A *sparse* autoencoder adds a penalty so that only a small number of coordinates of h are active for any given input. This discourages the network from simply memorizing inputs and instead encourages each hidden unit to capture a distinct, localized pattern in the data. The result is a set of learned features that tend to be more interpretable (Cunningham et al., 2023).

Movva et al. (2024)'s proposed approach is to: (1) train a sparse autoencoder on text embeddings to obtain a dictionary of interpretable features; (2) use regularized feature selection (see also Section 7.1) to identify the features that are most predictive of the outcome; and (3) repeatedly prompt an LLM with prototypical high- and low-activation texts to produce concise natural-language labels for those features. On real datasets, the method recovers patterns such as: surprise-themed headlines receive more clicks, and speeches mentioning illegal immigration are more likely to be delivered by Republicans.

While the specific hypotheses surfaced in these applications are more intuitive than surprising, the value of these approaches lies in making the creative step—i.e., mapping observed regularities into testable, natural-language hypotheses—more systematic and scalable. By contributing to these creative, heuristic stages of research, LLMs may complement more formal modeling and expand the toolkit available to experimental and theoretical economists.

7 Machine Learning and Model-Building Beyond Economics

This final section discusses related areas of work outside of economics: Section 7.1 considers the literature on interpretable machine learning; Section 7.2 asks whether large language models are merely predictive in a statistical sense, or if they have in some cases learned an underlying model.

7.1 Interpretable ML

The importance of interpretability in machine learning is powerfully illustrated by an incident described in Caruana et al. (2015). In a large-scale healthcare study, machine learning algorithms were trained to predict mortality risk for pneumonia patients. This prediction task had direct payoff consequences, since limited hospital capacity meant that risk assessment was critical for deciding which patients required admission and which should instead be treated as outpatients. Among the algorithms they considered, a neural network delivered the highest predictive accuracy. Yet it was ultimately rejected. Why?

One of the other algorithms the researchers trained was a more interpretable rule-based algorithm. When they examined this algorithm, they discovered that it had learned to classify patients with a history of asthma as low-risk—a counterintuitive and potentially dangerous classification. But this pattern did in fact appear in the data, because patients with asthma were typically admitted directly to intensive care, and received more aggressive treatment that succeeded in lowering their mortality rate. This relationship between asthma and mortality in the data is a classic example of correlation not causation, and in fact *reverses* the true causal effect of asthma on mortality risk. While this specific failure mode could have been corrected, it raised the concern that there were other, equally problematic statistical correlations in the data. The researchers thus chose to deploy a more transparent logistic regression model in place of the black box neural network, so that they could more directly evaluate what the algorithm was doing.

Episodes like this highlight the importance of understanding how a black-box model generates its predictions before entrusting it with important decisions. The field of interpretable machine learning (or "explainable ML") has developed a variety of techniques to open up the black box and examine its internal workings. These techniques can help determine whether a model has captured meaningful structure or is merely exploiting superficial correlations.

One widely used strategy is to assess which inputs have the greatest influence on a algorithm's predictions. For example, the Shapley value of a feature (Shapley, 1951; Lundberg and Lee, 2017) is its average impact on an algorithm's prediction. It is computed by comparing the algorithm's prediction with and without the feature across all choices for the remaining features, and averaging these changes in the algorithm's prediction. Intuitively, a feature is unimportant if the algorithm makes the same prediction with and without it, and the feature is important if the algorithm's predictions are responsive to the feature no matter what additional information it has. There are additionally notions of feature importance tailored to a specific model class. In LASSO regression, for instance, importance is often evaluated by the magnitude of the estimated coefficients, and for decision trees, importance of a given feature is measured by how much splitting on that feature reduces the outcome's entropy.

Another major family of methods for explaining model predictions relies on perturbing a model's inputs, internal representations, or training data (Koh and Liang, 2017), and observing how its predictions change. For example, in computer vision models, gradient-based saliency maps (Simonyan et al., 2014) identify which pixels most influence a prediction by computing the sensitivity of the class score to changes in each pixel (Zhou et al., 2016; Selvaraju et al., 2017). More advanced approaches operate on deeper feature maps but project their results back into the original space of pixels for interpretability (Zeiler and Fergus, 2014). Such approaches can highlight semantically meaningful regions—for example, revealing that a sofa, table, and fireplace are key to classifying a scene as a "living room " (Zhou et al., 2016).

Yet another class of approaches is to train a simpler, more interpretable model to approximate the behavior of a complex one. For example, Ribeiro et al. (2016) propose finding explanations for the prediction $f(x)$ at a particular input x by fitting an interpretable model in the neighborhood of x. Specifically, the analyst specifies a class of interpretable models (e.g., sparse linear models) and a measure of complexity for models in that class (e.g., the number of nonzero coefficients). The analyst then searches within that class for the model that best fits the black box locally around x, subject to a penalty on model complexity. The result is a low-complexity approximation of the black-box model's behavior around the point of interest. Similar ideas can be applied to learn a global surrogate model (such as a decision tree or rule list) that approximates the black box across the entire input space (Craven and Shavlik, 1996; Lakkaraju et al., 2016).

This literature has informed several of the approaches discussed earlier, including Hirasawa et al. (2022)'s use of feature-importance measures and Ludwig and Mullainathan (2024)'s use of gradient-based image morphing, among others. But its primary motivation is different from that of economic theory. In particular, much of the work on interpretable ML is concerned with whether an algorithm's predictions can be *trusted*. In many applications, it is acceptable for the algorithm to base predictions on a complex function of the inputs that exploits statistical correlations, provided those correlations are not misleading (as in the pneumonia example). In such cases, the aim is to characterize how the model uses its inputs and how its internal representations relate to its predictions. By contrast, economic modeling seeks to uncover the mechanisms that generate outcomes, with the aim of guiding interventions at the level of policies, institutions, or social structures.[15]

It is worth noting that interpretability and causality, while sometimes used interchangeably in economics, refer to distinct concepts and literatures in machine learning. The interpretable ML literature discussed above is concerned with explaining *how* a predictive model generates its outputs, regardless of whether those outputs have causal meaning. A variety of work in statistics and econometrics is more directly concerned with the causal interpretation of identified parameters; for overviews of this large topic from several perspectives see, e.g., Pearl (2009), Imbens and Wooldridge (2009), and Schölkopf et al. (2021). Economic modeling, as discussed in this article, ideally seeks models that both have legible structure and causal interpretations.

The next section concerns whether LLMs in particular have learned such structured representations.

7.2 "World Models"

Economics is not the only discipline in which models and explanations of underlying mechanisms are valued alongside predictive accuracy. The question of whether large language models (LLMs) have learned such models—rather than serving solely as high-performing predictors—has recently become an active topic of research.

A central issue in this broader discussion is whether LLMs can uncover what some researchers call a *world model*—a structured representation of the underlying rules or dynamics of a domain—or whether they are fundamentally limited to pattern-matching for prediction.

There is no consensus standard for how to evaluate whether the LLM possesses an understanding of a world model. One productive testbed has been to restrict attention to simple board games, where the "world state" is controllable and known. Specifically, suppose a large language model is given a large training set of play in the game, but is not given any information about the rules of the game. Will the model eventually come to uncover these rules and to track the evolution of the board?

Toshniwal et al. (2022) show that (when given sufficient data) a transformer is able to continue an initial sequence of chess moves with a valid following move. Li et al. (2023) replicate this finding for the game Othello, and moreover find evidence that the model has an internal representation of the state of the board. These papers

[15] The emerging field of mechanistic interpretability shares this broader ambition of developing a holistic understanding, though the economic models considered here are by design much simpler than the neural networks this literature seeks to explain.

provide evidence that the model is not simply memorizing a large set of statistical associations, but has in fact learned something about the game.

Vafa et al. (2024a) argue that predicting legal next moves is not sufficient evidence that a model has learned the game. They create a dataset consisting of taxi rides in New York City and train a foundation model on these sequences. They find that the trained model has strikingly good route planning abilities, usually identifying the shortest path between two points. But when Vafa et al. (2024a) reconstruct the model's implicit street map of New York City, they find that it has little resemblance to the actual map, and moreover contains streets with impossible physical orientations.

This distinction between predictive capability and an underlying world model is even starker in Vafa et al. (2025). They show that a foundation model can predict planetary orbits with near-perfect accuracy, yet fail to uncover the Newtonian laws that govern them. When fine-tuned on a small dataset where success would require such a model to apply Newtonian mechanics, it instead extrapolates poorly—suggesting it has learned surface patterns tailored to the training data, rather than the deeper physical principles that would generalize.

Taken together, these papers highlight two points. First, different ways of evaluating whether an LLM has uncovered a "world model" can lead to different conclusions. Second, for some reasonable definitions of a world model, high predictive performance does not, on its own, establish that a model has learned the correct underlying structure. Since the "world model" sought by these researchers closely parallels what this paper refers to as an "economic model," these findings further reinforce the continuing importance of theory.

8 Conclusion and Open Questions

Artificial intelligence is likely to change the practice of scientific research in significant ways over the coming years. This paper suggests that machine learning methods can be complementary to traditional structural economic modeling—not merely as predictive tools, but as tools for evaluating, improving, and extending economic theories. Alongside the application of these methods, there is a need for clearer understanding of what they do, the conditions under which they are most effective, and how to interpret their outputs.

Several open questions emerge from the work surveyed here.

On evaluating economic models. This paper has discussed several metrics for evaluating different dimensions of economic models, including their predictive completeness, restrictiveness, and portability across domains. These measures are not exhaustive. For example, the notion of portability introduced in Section 3.3 focuses on whether a model estimated in one domain predicts well in another, implicitly assuming that the same parameter values govern behavior across domains. But in many applications, a model's value lies precisely in its ability to identify a general structure that is shared across domains—such as risk aversion—while allowing its specific manifestation to vary, for instance through different parameter values. Such a model would have low transferability according to the measure in Section 3.3, but would be "generalizable" in another important sense. Distinguishing between undesirable flexibility and productive generality is therefore one example of a dimension of economic models that has not yet been clarified.

Quantifiable criteria for evaluating economic models will become increasingly important as theory-building itself becomes amenable to partial automation (see the next section). It is more challenging to optimize toward an objective that is not clearly specified and measured.

Beyond defining such metrics, an important research direction going forward will be systematic empirical assessment of models along these dimensions. For example, the papers in Section 3 evaluate models from behavioral and experimental economics, but we know relatively little about about the complex structural models that are commonly used in industrial organization and macroeconomics. It is not obvious ex ante how well these models predict out of sample relative to modern machine learning methods, nor how restrictive they truly are in practice.

On the role of AI in the economic research process. This paper has outlined several distinct ways in which machine learning methods may contribute to the development of economic models. One direction for this literature is fundamentally engineering-oriented: integrating these approaches into a more complete pipeline for theory generation. For example, a theorist might specify a candidate model, which an algorithm then stress-tests by identifying cases where it fails (as in Section 4.1). To then improve the theory for these cases, the resulting counterexamples could guide the generation of synthetic data by querying LLM "subjects" (as in Section 6.1). Finally, these data could be analyzed—using expert judgment, crowdsourcing, or automated tools—to identify systematic regularities that feed back into model refinement (as in

Section 4.2).

At the same time, progress along this frontier raises a parallel set of scientific questions about the use of AI within the research process. Section 6, for example, highlights fundamental questions around the use of LLMs as proxies for human decision-makers. What population, if any, does an LLM represent? To what extent can its behavior be reliably steered toward particular groups, contexts, or institutional settings? Most critically, will its responses generalize to experimental paradigms that were not represented in its training data? Addressing these questions requires principled methods for evaluating AI predictions and understanding when and how they diverge from reality.

Taken together, these considerations suggest that AI does not merely offer new tools for economists. Rather, it is a transformation that prompts a reconsideration of how economic knowledge is generated, evaluated, and justified.

References

Aher, G. V., R. I. Arriaga, and A. T. Kalai (2023): "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," in *Proceedings of the 40th International Conference on Machine Learning*, ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, PMLR, vol. 202 of *Proceedings of Machine Learning Research*, 337–371.

Andrews, I., D. Fudenberg, L. Lei, A. Liang, and C. Wu (2024): "The Transfer Performance of Economic Models," Working Paper.

Anthis, J. R., R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, E. Brynjolfsson, J. Evans, and M. S. Bernstein (2025): "Position: LLM Social Simulations Are a Promising Research Method," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, position Paper Track.

Athey, S. (2019): "The impact of machine learning on economics," in *The Economics of Artificial Intelligence: An Agenda*, ed. by A. Agrawal, J. Gans, and A. Goldfarb, Chicago: University of Chicago Press, 507–552.

Athey, S. and G. W. Imbens (2016): "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

——— (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

Ba, C., J. A. Bohren, and A. Imas (2023): "Over- and Underreaction to Information," Working Paper.

Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan (2022): "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*.

Bates, S., T. Hastie, and R. Tibshirani (2024): "Cross-validation: what does it estimate and how well does it do it?" *Journal of the American Statistical Asso-*

ciation, 119, 1434–1445.

Bengio, Y., A. Courville, and P. Vincent (2013): "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 1798–1828.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

Camerer, C. and T. Hua Ho (1999): "Experience-weighted Attraction Learning in Normal Form Games," *Econometrica*, 67, 827–874.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, 119, 861–898.

Camerer, C. F., G. Nave, and A. Smith (2019): "Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning," *Management Science*, 65, 1867–1890.

Caruana, R., Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad (2015): "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, KDD '15, 1721–1730.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017): "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, 107, 261–265.

Clark, C. and A. Storkey (2015): "Training deep convolutional neural networks to play go," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, JMLR, 1766–1774.

Compiani, G., I. Morozov, and S. Seiler (2025): "Demand Estimation with Text and Image Data," .

Cornia, M., L. Baraldi, G. Serra, and R. Cucchiara (2018): "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Transactions on Image Processing*, 27, 5142–5154.

Craven, M. and J. W. Shavlik (1996): "Extracting tree-structured representations of trained networks," in *Advances in Neural Information Processing Systems 8 (NIPS 1995)*, MIT Press, 24–30.

Crawford, V. P. and N. Iriberri (2007): "Fatal attraction: Salience, naivete, and sophistication in experimental 'hide-and-seek' games," *American Economic Review*, 97, 1731–1750.

Cunningham, H., A. Ewart, L. Riggs, R. Huben, and L. Sharkey (2023): "Sparse autoencoders find highly interpretable features in language models," *arXiv preprint arXiv:2309.08600*.

Ellis, K., S. Kariv, and E. Ozbay (2023a): "The Predictivity of Theories of Choice Under Uncertainty," Working Paper.

——— (2023b): "What can the demand analyst learn from machine learning?" Working Paper.

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep Learning for Individual Heterogeneity: An Automatic Inference Framework," .

Fava, B. (2025): "Training and Testing with Multiple Splits: A Central Limit Theorem for Split-Sample Estimators," *arXiv preprint arXiv:2511.04957*.

Fessler, P. and M. Kasy (2019): "How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions," *The Review of Economics and Statistics*, 101, 681–698.

Fudenberg, D., W. Gao, and A. Liang (2026): "How Flexible Is That Functional Form? Quantifying the Restrictiveness of Theories," *The Review of Economics and Statistics*, 108, 194–209.

Fudenberg, D. and G. Karreskog Rehbinder (2024): "Predicting cooperation with learning models," *American Economic Journal: Microeconomics*, 16, 1–32.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022): "Measuring the Completeness of Economic Models," *Journal of Political Economy*, 130, 956–990.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–4141.

Gao, W., S. Han, and A. Liang (2026): "How Well Do LLMs Predict Human Behavior? A Measure of Their Pretrained Knowledge," *arXiv preprint arXiv:2601.12343*, econometrics (econ.EM).

Gechter, M., C. Samii, R. Dehejia, and C. Pop-Eleches (2019): "Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference," .

Goldstein, W. M. and H. J. Einhorn (1987): "Expression theory and the preference reversal phenomena," *Psychological Review*, 94, 236–254.

Han, S. (2025): "Mining Causality: AI-Assisted Search for Instrumental Variables," School of Economics, University of Bristol.

Han, S. and K. Lee (2025): "Copyright and Competition: Estimating Supply and Demand with Unstructured Data," .

Hartford, J. S., J. R. Wright, and K. Leyton-Brown (2016): "Deep Learning for Predicting Human Strategic Behavior," in *Advances in Neural Information Processing Systems*, ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Curran Associates, Inc., vol. 29.

Hewitt, L., A. Ashokkumar, I. Ghezae, and R. Willer (2024): "Predicting Results of Social Science Experiments Using Large Language Models," Working paper.

Hirasawa, T., M. Kandori, and A. Matsushita (2022): "Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies," Working Paper.

Hornik, K., M. Stinchcombe, and H. White (1989): "Multilayer feedforward networks are universal approximators," *Neural Networks*, 2, 359–366.

Horton, J. J. (2023): "Large Language Models as Simulated Economic Agents: What Can We Learn from *Homo Silicus?*" NBER Working Paper 31122, National Bureau of Economic Research.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–291.

Koh, P. W. and P. Liang (2017): "Understanding Black-box Predictions via Influence Functions," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia: PMLR, 1885–1894.

Külpmann, P. and C. Kuzmics (2022): "Comparing theories of one-shot play out of treatment," *Journal of Economic Theory*, 205, 105554.

Lakkaraju, H., S. H. Bach, and J. Leskovec (2016): "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 1675–1684.

Lei, J. (2025): "A Modern Theory of Cross-Validation through the Lens of Stability," *arXiv preprint arXiv:2505.23592*.

Li, K., A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg (2023): "Emergent world representations: Exploring a sequence model trained on a synthetic task," in *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, X. and C. F. Camerer (2022): "Predictable Effects of Visual Salience in

Experimental Decisions and Games," *The Quarterly Journal of Economics*, 137, 1849–1900.

Lucas, R. E. (1980): "Methods and problems in business cycle theory," *Journal of Money, Credit and Banking*, 12, 696–715.

Ludwig, J. and S. Mullainathan (2024): "Machine Learning as a Tool for Hypothesis Generation*," *The Quarterly Journal of Economics*, 139, 751–827.

Ludwig, J., S. Mullainathan, and A. Rambachan (2025): "Large Language Models: An Applied Econometric Framework," .

Lundberg, S. M. and S.-I. Lee (2017): "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA.

McFadden, D. and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.

Mei, Q., Y. Xie, W. Yuan, and M. O. Jackson (2024): "A Turing test of whether AI chatbots are behaviorally similar to humans," *Proceedings of the National Academy of Sciences*, 121, e2313925121.

Milgram, S. (1963): "Behavioral study of obedience," *Journal of Abnormal and Social Psychology*, 67, 371–378.

Movva, R., K. Peng, N. Garg, J. Kleinberg, and E. Pierson (2024): "Sparse autoencoders for hypothesis generation," *arXiv preprint arXiv:2402.07857*.

Mullainathan, S. and A. Rambachan (2024): "From Predictive Algorithms to Automatic Generation of Anomalies," Working Paper.

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

Nagel, R. (1995): "Unraveling in guessing games: An experimental study," *American Economic Review*, 85, 1313–1326.

Pearl, J. (2009): "Causal Inference in Statistics: An Overview," *Statistics Surveys*, 3, 96–146.

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

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

Plonsky, O., I. Erev, T. Hazan, and M. Tennenholtz (2017): "Psychological forest: Predicting human behavior," *AAAI Conference on Artificial Intelligence*, 31, 656–662.

Raman, N., T. Lundy, T. Amin, J. Perla, and K. Leyton-Brown (2025): "STEER-ME: Assessing the Microeconomic Reasoning of Large Language Models," *arXiv preprint arXiv:2502.13119*.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016): ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, 1135–1144.

Rubinstein, A. (2006): "Dilemmas of an Economic Theorist," *Econometrica*, 74, 865–883.

Schelling, T. C. (1960): *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.

Schölkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio (2021): "Toward causal representation learning," *Proceedings of the IEEE*, 109, 612–634.

Selten, R. (1991): "Properties for a Measure of Predictive Success," *Mathematical Social Sciences*, 21, 153–167.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2017): "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shapley, L. S. (1951): "Notes on the n-Person Game – II: The Value of an n-Person Game," Tech. Rep. RM-670, RAND Corporation, Santa Monica, CA, aSTIA Document Number ATI 210720.

Simonyan, K., A. Vedaldi, and A. Zisserman (2014): "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, iCLR 2014 Workshop Submission.

Stahl, D. O. and P. W. Wilson (1994): "Experimental evidence on players' models of other players," *Journal of Economic Behavior and Organization*, 25, 309–327.

Toshniwal, S., S. Wiseman, K. Livescu, and K. Gimpel (2022): "Chess as a testbed for language model state tracking," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, Association for the Advancement of Artificial Intelligence, 11385–11393.

Vafa, K., P. G. Chang, A. Rambachan, and S. Mullainathan (2025): "What has a foundation model found? Using inductive bias to probe for world models," in *Proceedings of the 42nd International Conference on Machine Learning (ICML).*

Vafa, K., J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan (2024a): "Evaluating the world model implicit in a generative model," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS).*

Vafa, K., E. Palikot, T. Du, A. Kanodia, S. Athey, and D. M. Blei (2024b): "CAREER: A Foundation Model for Labor Sequence Data," *Transactions on Machine Learning Research*, creative Commons Attribution 4.0 International (CC BY 4.0).

Vapnik, V. N. and A. Y. Chervonenkis (1971): "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, 16, 264–280.

Vig, E., M. Dorr, and D. Cox (2014): "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2798–2805.

Wright, J. R. and K. Leyton-Brown (2014): "Level-0 meta-models for predicting human behavior in games," in *Proceedings of the fifteenth ACM conference on Economics and computation*, ACM, 857–874.

Zeiler, M. D. and R. Fergus (2014): "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, Springer, vol. 8689 of *Lecture Notes in Computer Science*, 818–833.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba (2016): "Learning Deep Features for Discriminative Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929.

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy (2021): "Domain generalization: A survey," Working paper.