Measuring the Completeness of Economic Models

Measuring the Completeness of Economic Models

Drew Fudenberg

Massachusetts Institute of Technology

Jon Kleinberg

Cornell University

Annie Liang

Northwestern University

Sendhil Mullainathan

University of Chicago

Economic models are evaluated by testing the correctness of their predictions. We suggest an additional measure, "completeness": the fraction of the predictable variation in the data that the model captures. We calculate the completeness of prominent models in three problems from experimental economics: assigning certainty equivalents to lotteries, predicting initial play in games, and predicting human generation of random sequences. The completeness measure reveals new insights about these models, including how much room there is for improving their predictions.

A two-page abstract of an early version of this project appeared in the Association of Computing Machinery Conference on Economics and Computation as "The Theory Is Predictive, But Is It Complete?" We thank Alberto Abadie, Amy Finkelstein, Indira Puri, Marciano Siniscalchi, Charlie Sprenger, and Johan Ugander for helpful comments. We

Electronically published February 25, 2022

Journal of Political Economy, volume 130, number 4, April 2022.

https://doi.org/10.1086/718371

I. Introduction

There is more reason to look for ways to improve a model that predicts poorly than one that predicts well. But what constitutes "good" performance? Our view is that the answer depends on how well the outcome could possibly be predicted given the specified "features" (i.e., the explanatory variables). To fix ideas, suppose we have data on whether customers agreed to a particular loan offer. Loan offers differ on characteristics such as the interest rate or the term of the loan. One model (the "NPV model") of how these characteristics relate to demand might posit that customers view loans through the lens of expected cost of capital over the duration of the loan. The expected cost of capital is a specific function of the available features. We could test the predictions of this model by evaluating it on data, for example, by seeing whether demand increases when the effective interest rate drops. These tests allow us to reject wrong models, but they do not tell us how much better a different model could do.

To get at this, we propose comparing the model's accuracy to that of the best prediction of demand that could be made using the features we have that describe each loan. Comparing the benchmark's predictive accuracy to that of the NPV model would tell us how much of the predictable signal in the outcome (given the baseline features) is captured by the NPV model. If the best predictions are much better than those of the NPV model, there may be another model built on the same features that substantially improves predictive accuracy. For example, another model might postulate that customers ignore future interest rates and focus only on the initial interest rate, or that 2.99% is viewed differently from 2.95%. On the other hand, if the best predictions are not much better than those of the NPV model, then alternative models built on the same features cannot possibly do much better on this data set. For new models to help, they must identify new variables that are not currently measured. For example, models that emphasize framing and persuasion would point to expanding our data set to include the vocabulary used in the loan descriptions.

Moving beyond this specific example, any model's prediction error can generally be decomposed into two sources: (1) intrinsic noise in the outcome due to limitations of the features we have measured, that is, the irreducible error, and (2) regularities in the data that the model does not capture. The irreducible error provides an upper bound for how well any model (based on the measured features) could possibly do.

are also grateful to Adrian Bruhin, Helga Fehr-Duda, Thomas Epper, Kevin Leyton-Brown, and James Wright for sharing data with us, and we are grateful for financial support from National Science Foundation grants SES 1643517, 1851629, and 195105. Data are provided as supplementary material online. This paper was edited by Emir Kamenica.

A benchmark at the other end is the performance of a baseline model, such as "guess the outcome at random."[1] We use these extremes to measure what we call the "completeness" of a model:

$$\frac{\mathcal{E}_{\text{base}} - \mathcal{E}_{\text{model}}}{\mathcal{E}_{\text{base}} - \mathcal{E}_{\text{irreducible}}},$$

where $\mathcal{E}_{\text{base}}$ is the out-of-sample prediction error under the baseline, $\mathcal{E}_{\text{model}}$ is the out-of-sample error of the model, and $\mathcal{E}_{\text{irreducible}}$ is the irreducible error. That is, completeness is the model's reduction in prediction error relative to the baseline, divided by the achievable reduction in prediction error. A model with a completeness of 0 does not improve upon the baseline, while a model with a completeness of 1 eliminates all but the irreducible error. Crucially, a model can be complete for the given measured features even if it predicts poorly and its R^2 is low. The distinction between a complete model and a perfectly predictive model is especially relevant for the social sciences, where we expect there to be substantial irreducible noise in most outcomes of interest given the measurable features. Economists can rarely hope for our models to be perfectly predictive, but we can hope for them to be relatively complete.

In addition to proposing the completeness measure, we demonstrate that completeness can be precisely estimated for a diverse range of experimental data sets. The challenge is estimating irreducible error. In general, the performance of black-box machine-learning methods can be used as a stand-in for irreducible error. When the data consist of a large number of outcome observations for each vector of features, then it is possible to obtain a fairly precise estimate of irreducible error using a simple "lookup table," which nonparametrically searches the space of possible models, and finds the model that maximizes out-of-sample predictive accuracy for the set of available features. Many lab data sets have this property; for example, the data sets may contain a large number of observations of game play for each of a small set of games, or a large number of observations of certainty equivalents for each of a small set of lotteries.

Our applications use data sets like this to evaluate the completeness of prominent models from three experimental domains: cumulative prospect theory (Tversky and Kahneman 1992) for prediction of certainty equivalents, the Poisson cognitive hierarchy model (Camerer, Ho, and Chong 2004) for prediction of initial play in games, and Rabin and Vayanos (2010) for prediction of human perception of randomness.

These applications illustrate three general points. First, the absolute error of a model can be extremely misleading. Cumulative prospect theory

[1] Diebold and Kilian (2001) propose benchmarking the accuracy of time series forecasts relative to that of a bad forecast. This is in the spirit of our comparison against a baseline model.

has very poor absolute fit (a mean-squared error of 67.78), but it is 94% complete: Its high error rate is nearly irreducible given the available features. Second, a model's absolute gain over the baseline can be equally misleading. For example, the Rabin and Vayanos (2010) model reduces prediction error relative to the baseline by only 0.0006, but is nevertheless 10% complete. Finally, the completeness measure reveals the relative complexity of various prediction problems. For example, in our initial play application we find that the completeness of the Poisson cognitive hierarchy model differs substantially across classes of 3×3 games, varying from 68% to 97%. This suggests important underlying differences between games that need to be better understood. These, and the subsequent observations we make in sections V.A–V.D, are informative about the problem domains and how much room there is for improving the predictions of their leading models without obtaining new sorts of data.

Our completeness measure depends on a specified set of features and is evaluated for a given prediction problem. If we change the feature set or the data, we would expect the measurement of completeness to change, as we discuss in section VI.B. Likewise, the completeness of a model depends on the specified prediction problem: With the same features, a model of the effect of a price cut on sales might be able to predict the aggregate effect (e.g., a 5% increase in sales) very well but be unable to predict which consumers would increase their purchases.

Related work.—Irreducible error is an old concept in statistics and machine learning. A large literature has studied the decomposition of this error into *bias* (reflecting error due to the specification of the model class) and *variance* (reflecting sensitivity of the estimated rule to the randomness in the training data). Depending on the quantity of data available to the analyst, it may be preferable to trade off bias for variance or vice versa. This paper abstracts from these concerns, as well as the related concern of overfitting. We work exclusively with data sets in which there are enough data that the best feasible out-of-sample prediction accuracy is well approximated by searching across the unrestricted space of mappings from features into outcomes (see app. A).

The only previous measures of predictive success for economic models in experimental work that we know of are Selten's (1991) measure of the relative frequency of successful predictions, Erev et al.'s (2007) definition of the *equivalent number of observations*, and Apesteguia and Ballester's (2021) measure of goodness-of-fit for stochastic choice models. Our work differs in that we focus on understanding the best possible prediction in a given problem, and evaluate performance relative to that benchmark.

Several recent papers compare a model's predictive performance to that of specific machine-learning algorithms. These algorithms sometimes approximate the best possible predictions. For example, Peysakhovich and Naecker (2017) compare the performance of economic models of

the willingness to pay for three-outcome lotteries to the performance of regularized regression algorithms, and Bodoh-Creed, Boenhke, and Hickman (2019) compare the performance of simple OLS models using known regressors against the performance of random forests built on a rich feature set, for the problem of predicting pricing variation. The algorithms used in these papers need not achieve the irreducible error, but they do provide a lower bound for the best achievable accuracy. We show that in experimental contexts, it can be possible to directly estimate the best achievable accuracy and use that as a benchmark.

Other papers directly use an algorithmic approach to predict economic behavior, for example, Noti et al. (2016), Plonsky et al. (2017, 2019), and Zhao et al. (2020) for prediction of choice, and Camerer, Nave, and Smith (2019) for prediction of disagreements in bargaining. The improvements achieved by these more complex algorithms over the existing economic models are sometimes modest. One reason for this might be intrinsic noise, as Bourgin et al. (2019) point out. We show how this noise can be quantified.

Finally, we note that in the special case in which performance is measured by mean-squared error and the baseline is an unconditional mean, our completeness measure can be seen as a ratio of the model's R^2 and the nonparametric R^2, as we explain in appendix B. Our approach is not special to this loss function, however, and can be implemented with any metric of accuracy.

II. Example

We begin with a simple example that illustrates the need for a measure such as completeness. Let $y \in \{0, 1\}$ be a binary outcome of interest, which is related to two binary features x_1 and x_2, each of which has an independent probability 0.5 of taking the value 1. Specific theories make predictions about how the given features relate to the outcome. Suppose that our model posits that the features enter linearly according to $y = \beta(x_1 + x_2)$ for some $\beta \in \mathbb{R}$. We can test this model by acquiring observations of (x_1, x_2, y) drawn from their true joint distribution, estimating β, and using the estimated model to predict outcomes in a new data set. The performance of the model is evaluated according to some loss function, for example, the (average) squared difference between the prediction $\hat{y}$ and the true outcome y, that is, $-(y - \hat{y})^2$. But it is hard to interpret the magnitude of this error without additional information. To see the problem, consider table 1, which describes two data-generating processes for y given x_1 and x_2.

For both data-generating processes, the estimated value of the parameter β (given sufficient data) is $\beta = 0.5$, so the estimated model is $f(x) = 0.5(x_1 + x_2)$. The expected mean-squared error of this model is

TABLE 1
TWO PROCESSES SPECIFYING THE EXPECTED VALUE OF y GIVEN THE VALUES OF x_1 AND x_2

x_1	x_2	$\mathbb{P}(y = 1 \mid x)$
Process 1:		
0	0	0
0	1	.5
1	0	.5
1	1	1
Process 2:		
0	0	0
0	1	.1
1	0	.9
1	1	1

NOTE.—In both cases, the distribution over features is uniform.

0.125 under both of the data-generating processes in table 1.[2] Taking the magnitude of the prediction error at face value would suggest that the model is equally predictive for both versions of the ground truth. But the equality of the prediction errors obscures an important difference, which is that in the first case there is no alternative model built on x_1 and x_2 that can make more accurate predictions, while in the second case the model $y = \beta_1 x_1 + \beta_2 x_2$ (with β_1 estimated to 0.1 and β_2 estimated to 0.9) achieves a prediction error of 0.045. That is, the proposed model is complete given the first data-generating process but incomplete given the second. Our subsequent approach formalizes this notion of completeness.

III. Completeness

Section III.A introduces the setting of prediction problems and section III.B defines completeness.

A. *Preliminaries*

In a prediction problem, there is an outcome Y whose realization is of interest, and features X that are statistically related to the outcome. The goal is to predict the outcome given the observed features. Some examples include predicting an individual's future wage based on childhood covariates (city of birth, family income, quality of education, etc.), or predicting a criminal defendant's flight risk based on the defendant's past record

[2] For both data-generating processes, the prediction given $x = (0, 0)$ is $\hat{y} = 0$, and the prediction given $x = (1, 1)$ is $\hat{y} = 1$. Both result in a conditional error of zero. The model predicts $\hat{y} = 0.5$ for the vectors $x = (0, 1)$ and $x = (1, 0)$, which has a conditional error of $1/4$ (under both data-generating processes). Thus the unconditional error is $(1/2)(1/4) = 1/8$.

and properties of the crime (Kleinberg et al. 2018). We focus on three prediction problems that emerge from experimental economics:

EXAMPLE 1 (risk preferences). Can we predict the valuations that people will assign to various money lotteries?

EXAMPLE 2 (predicting play in games). Can we predict how people will play the first time they encounter a new simultaneous-move game?

EXAMPLE 3 (human generation of random sequences). Given a target random process—for example, a Bernoulli random sequence—can we predict the errors that a human will make while mimicking this process?

Formally, suppose that the observable features belong to some space $\mathcal{X}$ and the outcome belongs to $\mathcal{Y}$. There is a true but unknown joint distribution P over $\mathcal{X} \times \mathcal{Y}$. A map $f : \mathcal{X} \to \mathcal{Y}$ from features to outcomes is a *prediction rule*. Many economic models can be described as a family of prediction rules or "models" $\mathcal{F}_\Theta$ indexed by an interpretable parameter set Θ. For example, the model class may impose a linear relationship $f(x) = \langle x, 0 \rangle$ between the outcome and a set of features x, in which case the parameter $\theta \in \Theta$ defines a vector of weights applied to each feature. In section V.A, one specification of $\mathcal{F}_\Theta$ is the family of certainty equivalents under utility functions $u(z) = z^\theta$ over dollar amounts, where the parameter θ reflects the degree of risk aversion.

B. Definition

We suppose that our prediction problem comes with a *loss function*, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $\ell(y', y)$ is the error assigned to a prediction of y' when the realized outcome is y. The commonly used loss functions mean-squared error and classification error correspond to $\ell(y', y) = (y' - y)^2$ and $\ell(y', y) = \mathbf{1}(y' \neq y)$, respectively.[3]

DEFINITION 1. The *expected error* (or *risk*) of prediction rule f on a new observation $(x, y) \sim P$ is

$$\mathcal{E}_P(f) = \mathbb{E}_P[\ell(f(x), y)]. \tag{1}$$

Let

$$f^*_\Theta \in \arg\min_{f \in \mathcal{F}_\Theta} \mathcal{E}_P(f)$$

denote any prediction rule in the parametric class $\mathcal{F}_\Theta$ that minimizes the expected prediction error. The expected error of any such "best" rule in $\mathcal{F}_\Theta$ is $\mathcal{E}_P(f^*_\Theta)$. In section IV, we discuss how to estimate $\mathcal{E}_P(f^*_\Theta)$ on finite data; here we discuss how to interpret it.

[3] Different loss functions are typically used when predicting distributions; see, e.g., Gneiting and Raftery (2007).

To understand a model's error, it is helpful to distinguish between two different sources of error. First, if the conditional distribution $Y|X$ is not degenerate, then even an ideal prediction rule

$$f^*(x) \in \arg\min_{y' \in \mathcal{Y}} \mathbb{E}_P[\ell(y', y)|x]$$

does not predict perfectly.

DEFINITION 2. The *irreducible error* in the prediction problem is the expected error

$$\mathcal{E}_P(f^*) = \mathbb{E}_P[\ell(f^*(x), y)] \tag{2}$$

of the ideal rule on a new test observation.

The irreducible error is a lower bound on the error when predicting Y using the features in $\mathcal{X}$.

A second source of prediction error is the specification of the class $\mathcal{F}_\Theta$. Typically, the best possible model will not be an element of $\mathcal{F}_\Theta$, as most model classes are at least slightly misspecified. If $\mathcal{F}_\Theta$ leaves out an important regularity in the data, then there may be models outside of $\mathcal{F}_\Theta$ that yield much better predictions.[4]

These two sources of prediction error have very different implications for how to generate better predictions. If the model's prediction error is substantially higher than the irreducible error, it may be possible to identify new regularities and incorporate them into models that improve prediction given the same feature set. These new models might be preferable if they do not involve too great an increase in complexity or in the number of parameters. Conversely, if the model's prediction error is close to the irreducible error for the current feature set, the priority should be to identify additional features that will allow for better predictions.

We propose the ratio of the reduction in prediction error achieved by the model, compared to the achievable reduction, as a measure of how close the model comes to the best achievable performance. We call this ratio the model's *completeness*. To operationalize this measure, we select a (potentially randomized) baseline $f_{\text{base}} \in \Delta(\mathcal{F}_\Theta)$ suited to the prediction problem. For example, in the prediction of certainty equivalents, prediction of the lottery's expected value is a natural baseline. The performance of this baseline rule is interpreted as a "worst case" prediction accuracy.

[4] On the other hand, expanding the model class risks overfitting, so more parsimonious model classes can lead to more accurate predictions when data are scarce (Hastie, Tibshirani, and Friedman 2009). As we discuss under "related work" in sec. I and in sec. IV, all of the data sets that we consider here are large relative to the number of features.

DEFINITION 3. The *completeness* of model class $\mathcal{F}_\Theta$ is

$$\frac{\mathcal{E}_P(f_{\text{base}}) - \mathcal{E}_P(f_\Theta^*)}{\mathcal{E}_P(f_{\text{base}}) - \mathcal{E}_P(f^*)}. \tag{3}$$

Completeness is a normalized measure of the reduction in error. A model with completeness 0 does no better than the baseline, while a "fully complete" model with completeness 1 removes all but the irreducible error.[5]

C. Discussion

1. Choice of Baseline

The proposed measure uses a baseline f_{base} to evaluate a given model's error.[6] For example, in our application to lotteries, the fact that the mean-squared error of the expected value is approximately 103 provides a useful comparison for the cumulative prospect theory (CPT) error of 68. Similarly, for predicting play in games, one should not expect any model to do worse than guessing at random, which leads to a misclassification rate of 1/3. In many cases, there is a natural choice for the baseline, or a range of natural choices. In appendix C1, we expand one of our applications by estimating completeness relative to a set of possible baselines. We show that completeness is stable across these choices.

An alternative to a user-specified baseline would be to use the best unconditional prediction of Y, for example, the average value of Y if the loss function is mean-squared error. The same rule would then be used across applications, which has the advantage of eliminating flexibility (and potential arbitrariness) in the choice of baseline. However, this fixed rule has some disadvantages. First, the unconditional prediction baseline may not be in the class of models being considered, and as a result, it can even yield negative completeness.[7] Second, the performance of the unconditional prediction baseline is very sensitive to the variability of the elements in $\mathcal{X}$ and can in principle be very large. As we show in appendix C1, the error of this empirical baseline is an order of magnitude larger than the other errors we report in our first application.

[5] This is one of many possible measures with completeness 1 when the model removes all but irreducible error and 0 when the model coincides with the baseline error. Our definition measures "units" of completeness as percentage improvements in prediction error, which facilitates comparison across settings with different loss functions.

[6] This is analogous to the importance of the choice of the null hypothesis in classical hypothesis testing: different choices of the null can lead to different conclusions from the same data.

[7] Also note that this measure has some unfortunate small-sample properties; e.g., if the data set consists of a single observation, then mechanically this empirical achieves the best possible error, so the denominator of eq. (3) is 0 and completeness is either undefined or equal to $-\infty$.

2. Dependence on Domain $\mathcal{X}$

The measure also depends on the underlying feature space X, that is, the set of lotteries or games in the experimental data set.[8] We show in section V.C that the completeness of the Poisson cognitive hierarchy model varies depending on which set of games is in the data set. This reflects the fact that some models perform better for certain kinds of inputs (e.g., certain kinds of lotteries or certain kinds of games). We view the ability of the completeness measure to capture this variation as a strength of the approach.

3. Expanding $\mathcal{X}$

Completeness is defined for a fixed feature set $\mathcal{X}$, which we generally interpret as the measured features in the data. If we vary $\mathcal{X}$ by expanding it to include new measured features, then the predictive performance of the original model remains the same, but the predictive optimum weakly improves. So a model that is complete for one feature set $\mathcal{X}$ may not be complete for another $\mathcal{X}' \supset \mathcal{X}$. In general, if a model is nearly complete for the measured features, the only way to improve predictive accuracy is to measure new features and develop new models on the larger feature set.

D. Evaluating Models

Predictive accuracy is only one of many criteria that matter for selecting theories. Economists typically also value parsimony, portability, and causal explanations, and trade them off against accuracy and each other when selecting models.[9] This paper is not designed to be about the tension between these criteria. Rather, our definition of completeness is meant as a tool to facilitate making such trade-offs.

A high-level analogy is to the idea of polynomial-time approximation algorithms for NP-hard optimization problems. In the theory of approximation algorithms, there is a tension between efficient algorithms (which run quickly and produce suboptimal solutions in general) and the optimal solution (which may be hard to find, but whose value cannot be improved). To even state this tension, one needs the notion of "the optimal solution" in the first place. This is obvious in the context of optimization problems and efficient algorithms for them, but as far as we know, the analog of the "optimal solution" is not in common use in experimental

[8] For naturally occurring data, the elements of $\mathcal{X}$ may be governed by an external process. In experimental data, they are generally chosen by the experimentalist.

[9] See Gabaix and Laibson (2008) and Hofman et al. (2021) for perspective pieces that articulate various facets of economic modeling, and Athey (2019) for a survey describing the opportunities created by machine learning for some of these goals.

settings. Evaluating the completeness of a model makes it possible to talk about the tension between simple and complex models, as well as related trade-offs, with reference to how close these models come to the best achievable performance.

Restrictive versus complete models.—In general, the more flexible a model is, the higher its completeness. At the extreme, a model class $\mathcal{F}_\Theta$ that includes all possible mappings from $\mathcal{X}$ to $\mathcal{Y}$ achieves full completeness. But such a model is also vacuous, as it has no falsifiable predictions. For a fixed level of predictive accuracy, we thus prefer models that are more restrictive. Fudenberg, Gao, and Liang (2020) provide an algorithmic measure of a model's "restrictiveness" by evaluating the completeness of the model on a range of synthetic data. Since the best achievable error varies substantially across data sets, low absolute error on these data sets is not enough to conclude that a model is unrestrictive; likewise, high absolute error does not imply that the model imposes substantial restrictions. But a model that is complete on all data is not restrictive.

Interpretable versus predictive models.—In many applications, researchers may prefer to sacrifice some predictive power and completeness to use a model that is easier to interpret, for example, using a model of preferences to predict choice as opposed to a black box. Having a measure for completeness tells us how much we sacrifice in terms of predictive power by requiring the model to be interpretable. In some cases, such as the CPT model in section V.A, it turns out that simple and interpretable models achieve completeness comparable to that of black-box algorithms, meaning this trade-off is not present.[10] On the other hand, the Rabin and Vayanos (2010) model achieves only partial completeness; this could be because a better, interpretable model exists, or it could be because human behavior in this domain is fundamentally complex and cannot be captured by a simple model. Having the measure of completeness makes it possible to describe this trade-off.

IV. Estimating Completeness from Finite Data

We now discuss how to estimate completeness from finite data. When the feature space $\mathcal{X}$ is "small" and there are a large number of observations of Y for each unique $x \in \mathcal{X}$, then a natural estimator for the irreducible error is the performance of a lookup table, which simply learns the best prediction of Y for each x. While the assumption that there is a large number of observations per x may seem demanding, it turns out to be satisfied for a potentially large number of experimental data sets, including the ones we subsequently study. We view a substantial part of the contribution of

[10] Fudenberg and Liang (2019) demonstrate a similar point for the domain of initial play in matrix games.

this paper as demonstrating that irreducible error can be approximated simply across a diverse range of experimental contexts.

When nonparametric estimation of irreducible error via a lookup table is not feasible, then econometric methods such as splines, sieves, and black-box machine-learning algorithms (e.g., lasso regression) can potentially be used as substitutes.[11] In those cases, the estimate of the ratio of the model's improvement relative to the improvement achieved by the black box can be seen as an upper bound on the completeness of the model, as in Peysakhovich and Naecker (2017).

We subsequently describe in detail the estimators we use in this paper, with section IV.A describing our estimators for the expected prediction errors in (3), and section IV.B describing our estimator for completeness.

A. *Estimators for Expected Prediction Errors*

Our approach applies to an arbitrary set $\mathcal{F}$ of maps from $\mathcal{X}$ to $\mathcal{Y}$. The special cases $\mathcal{F} = \{f_{\text{base}}\}$, $\mathcal{F} = \mathcal{F}_\Theta$, and $\mathcal{F} = \mathcal{X}^\mathcal{Y}$ (i.e., the unrestricted set of all possible maps from features in $\mathcal{X}$ into outcomes in $\mathcal{Y}$) respectively return the desired prediction errors $\mathcal{E}_P(f_{\text{base}})$, $\mathcal{E}_P(f_\Theta^*)$, and $\mathcal{E}_P(f^*)$ from (3).

In each case, we select a mapping from $\mathcal{F}$ based on a set of training observations, and evaluate the out-of-sample prediction error of the chosen mapping. Our estimator for the expected prediction error is the 10-fold cross-validated out-of-sample error. We describe this procedure in some detail as it may be new for some readers, but it is standard, and familiar readers may skip directly to section V.

1. Split data into $K = 10$ folds. All of the available data are randomly split into K equally sized disjoint subsets $Z_1, \ldots, Z_K$. In each iteration $1 \leq i \leq K$ of the procedure, the subset $Z^i_{\text{test}} \equiv Z_i$ is identified as the test data and the remaining subsets $Z^i_{\text{train}} \equiv \cup_{j \neq i} Z_j$ are used as training data.
2. Select a mapping from $\mathcal{F}$ that best fits the training data. For each iteration $i \in \{1, \ldots, K\}$ and mapping f, the in-sample performance of f for predicting the observations in Z^i_{train} is

$$e(f, Z^i_{\text{train}}) = \frac{1}{|Z^i_{\text{train}}|} \sum_{(x,y) \in Z^i_{\text{train}}} \ell(f(x), y)).$$

   This is a sample analog of the expected prediction error in (1). Choose any $f_i \in \arg\min_{f \in \mathcal{F}} e(f, Z^i_{\text{train}})$.[12]

[11] These methods may have better finite-sample performance when suitable regularity assumptions apply, but those assumptions may not be directly testable.

[12] When there are multiple minimizers, choose between them randomly.

3. Evaluate how well the chosen mapping performs out of sample. The selected model f_i is subsequently evaluated on the set of test observations in Z_{test}, where this model's out-of-sample performance on the ith test set is

$$\text{CV}_i = e(f_i, Z^i_{\text{test}}). \tag{4}$$

4. Average over out-of-sample errors. The average out-of-sample error across the K test sets is

$$\text{CV}\left(\mathcal{F}, \{Z_i\}_{i=1}^K\right) = \frac{1}{K}\sum_{i=1}^K \text{CV}_i. \tag{5}$$

B. *Estimator for Completeness*

Define

$$\begin{aligned}
\hat{\mathcal{E}}_{\text{base}} &\equiv \text{CV}\left(\{f_{\text{base}}\}, \{Z_i\}_{i=1}^K\right),\\
\hat{\mathcal{E}}_{\Theta} &\equiv \text{CV}\left(F_{\Theta}, \{Z_i\}_{i=1}^K\right),\\
\hat{\mathcal{E}}_{\text{best}} &\equiv \text{CV}\left(X^{\mathcal{Y}}, \{Z_i\}_{i=1}^K\right).
\end{aligned}$$

Subsequently, we refer to these estimates simply as prediction errors, understanding that they are finite-data estimates. In place of the theoretical completeness measure described in (3), we compute the empirical ratio

$$\frac{\hat{\mathcal{E}}_{\text{base}} - \hat{\mathcal{E}}_{\Theta}}{\hat{\mathcal{E}}_{\text{base}} - \hat{\mathcal{E}}_{\text{best}}} \tag{6}$$

from our data. The tables that we report in the subsequent applications in sections V.A–V.D are structured as in table 2.

Theoretical guarantees.—The empirical quantities $\hat{\mathcal{E}}_{\text{base}}$, $\hat{\mathcal{E}}_{\Theta}$, and $\hat{\mathcal{E}}_{\text{best}}$ are consistent estimators for $\mathcal{E}_P(f_{\text{base}})$, $\mathcal{E}_P(f^*_{\Theta})$, and $\mathcal{E}_P(f^*)$, respectively (Hastie, Tibshirani, and Friedman 2009), and the empirical estimate of completeness in (6) is a consistent estimator for (3).

These estimates are good approximations for the theoretical quantities when the number of observations is sufficiently large. In particular, for

TABLE 2
OUR RESULTS IN THE SUBSEQUENT APPLICATIONS

	Error	Completeness (%)
Baseline	$\hat{\mathcal{E}}_{\text{base}}$	0
Economic model	$\hat{\mathcal{E}}_{\Theta}$	$100 \times (\hat{\mathcal{E}}_{\text{base}} - \hat{\mathcal{E}}_{\Theta})/(\hat{\mathcal{E}}_{\text{base}} - \hat{\mathcal{E}}_{\text{best}})$
Irreducible error	$\hat{\mathcal{E}}_{\text{best}}$	100

$\hat{\mathcal{E}}_{\text{best}}$ to be a good approximation of the irreducible noise $\mathcal{E}_P(f^*)$, the analyst must have access to a sufficiently large number of observations for each distinct $x \in \mathcal{X}$. This can be a demanding criterion. To evaluate whether we have "enough" data in our applications, we report bootstrapped standard errors (with 1,000 draws) for our estimates.[13]

We additionally consider two tests in appendixes A1 and A2. First, we compare the performance of the lookup table with a machine-learning algorithm that is better suited to smaller data sets (bagged decision trees). The out-of-sample performances are comparable, but the lookup table has a lower error for all of our applications (see app. A1). Second, we investigate whether the out-of-sample performance of the lookup table has converged by evaluating its performance on subsamples of our data. The prediction errors using just 70% of the data are very close to those using all of our data. These analyses suggest that our estimate for irreducible error is a reasonable approximation in each of our applications.

In general, the condition that the data include many observations per feature is easier to satisfy in experimental settings, where the experimentalist has control over the structure of the data and can choose to acquire a large number of observations for each of a fixed set of feature values.[14]

V. Three Applications

A. *Application 1: Assigning Certain Equivalents to Lotteries*

1. Background and Data

An important question in economics is how individuals evaluate risk. In addition to expected-utility models (von Neumann and Morgenstern 1944; Samuelson 1952; Savage 1954), one of the most influential models of decision making under risk is cumulative prospect theory (Tversky and Kahneman 1992). This model provides a flexible family of risk preferences

[13] For applications 1 and 3, we report standard errors using a block bootstrapping procedure that clusters together all observations from the same subjects. Specifically, when generating a bootstrap sample, we randomly sample from the set of unique subjects with replacement, and include all observations associated with these subjects. We then carry out our (cross-validated) estimation of completeness on each bootstrap sample. Since we do not have complete subject ID data for application 2, we instead generate bootstrap samples by sampling from the set of all observations with replacement. Also, we also report standard errors for completeness using the analytic standard errors reported in Fudenberg, Gao, and Liang (2020); see app. A3. These standard errors are quite comparable to the bootstrapped standard errors that we report in the main text.

[14] In the data sets that we consider, there is an average of 179 observations per unique x for estimation of a mean (sec. V.A), 50 observations per unique x for estimation of a mode (most likely) outcome (sec. V.C), and 164 observations per unique x for estimation of a mean (sec. V.D).

that accommodates various behavioral anomalies, including reference-dependent preferences and nonlinear probability weighting.

A standard experimental paradigm for eliciting risk preferences, and thus for evaluating these models, is to ask subjects to report certainty equivalents for lotteries—that is, the lowest certain payment that the individual would prefer over the lottery. We consider a data set from Bruhin, Fehr-Duda, and Epper (2010), which includes 8,906 certainty equivalents elicited from 179 subjects, all of whom were students at the University of Zurich or the Swiss Federal Institute of Technology Zurich. Subjects reported certainty equivalents for the same 50 two-outcome lotteries, half over positive outcomes (e.g., gains) and half over negative outcomes (e.g., losses).

2. Prediction Task and Models

In this data set, the outcomes are the reported certainty equivalents for a given lottery, and the features are the lottery's two possible monetary prizes $\bar{z} > \underline{z}$ and the probability p of the first prize. A prediction rule is any function that maps the tuple $(\bar{z}, \underline{z}, p)$ into a prediction for the certainty equivalent. We use mean-squared error as the loss function: In a test set of n observations $\{(\bar{z}_i, \underline{z}_i, p_i; y_i)\}_{i=1}^{n}$—where $(\bar{z}_i, \underline{z}_i, p_i)$ is the lottery shown in observation i and y_i is the reported certainty equivalent—the mean-squared error of f is

$$\frac{1}{n}\sum_{i=1}^{n}[f(\bar{z}_i, \underline{z}_i, p_i) - y_i]^2.$$

We evaluate a prediction rule based on cumulative prospect theory (CPT),[15] which predicts

$$v^{-1}\big(w(p)v(\bar{z}) + [1 - w(p)]v\big(\underline{z}\big)\big)$$

for each lottery, where w is a probability weighting function, v is a value function, and by convention $|\bar{z}| > |\underline{z}|$. We follow Bruhin, Fehr-Duda, and Epper (2010) in our choice of functional forms:

$$v(z) = \begin{cases} z^{\alpha} & \text{if } z > 0, \\ -\big(-z^{\beta}\big) & \text{if } z \leq 0, \end{cases} \qquad w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1 - p)^{\gamma}}. \tag{7}$$

This model has four free parameters: $\alpha, \beta, \delta, \gamma \in \mathbb{R}_{+}$.

Finally, as a baseline, we predict the expected value of the lottery, which is $p\bar{z} + (1 - p)\underline{z}$. As we report in appendix C1, completeness of CPT is very similar to other benchmarks such as risk-averse variants of expected utility.

[15] CPT and the original prospect theory are equivalent on the two-outcome lotteries that we consider.

TABLE 3
CPT PREDICTS BETTER THAN EXPECTED VALUE

	Error
Baseline	104.63
	(10.14)
CPT	67.78
	(8.37)

3. Results

Table 3 reveals that CPT's out-of-sample predictions improve upon the expected-value benchmark.[16] CPT does much better than the expected-value benchmark, but falls far short of perfect prediction. It is difficult to interpret the size of CPT's error based on table 3 alone. It is not surprising that these models do not achieve perfect prediction, as we expect different subjects to report different certainty equivalents for the same lottery, and thus a model that provides the same prediction for each $(\bar{z}, \underline{z}, p)$ input cannot possibly predict every reported certainty equivalent. But besides the intrinsic variation in certainty equivalents for any fixed lottery, another potential source of error is the functional form imposed in (7). Could a different (potentially more complex) specification for the value function or probability weighting function lead to large gains in prediction? Relatedly, might there be other features of risk evaluation, yet unmodeled, which lead to even larger improvements in prediction?

To separate these sources of error, we need to understand how CPT's error compares to the irreducible error for these data. We estimate the irreducible error in this problem using a lookup table, where each of the 50 unique lotteries is mapped to the average certainty equivalent for that lottery in the training data. With 179 observations for each of the lotteries, we are able to approximate the mean certainty equivalent for each lottery using the training data, thus (approximately) minimizing the out-of-sample prediction error. We report the estimated irreducible error and its standard error in table 4.

Table 4 shows that the CPT prediction error is almost as low as the irreducible error; CPT achieves 94% of the feasible reduction in prediction error over the baseline.[17] Thus these data suggest that no theory that uses only the features $(\bar{z}, \underline{z}, p)$ can predict much better than CPT.[18] To further reduce error, we would need to expand the set of variables on which the

[16] The parameter estimates for CPT are $\alpha = 0.8$, $\beta = 1.2$, $\delta = 0.9$, and $\gamma = 0.5$.

[17] In app. C2, we show that completeness is nearly identical for other popular functional form specifications of CPT.

[18] It is hard to know whether the high completeness of CPT (in the specified functional form) comes from its good match to actual behavior or is because it is flexible enough to mimic most functions in $\mathcal{X}^{\mathcal{Y}}$. This question is explored in Fudenberg, Gao, and Liang (2020).

TABLE 4
CPT Is Nearly Complete for Prediction of Our Data

	Error	Completeness (%)
Baseline	104.63	0
	(10.14)	
CPT	67.78	94
	(8.37)	(2.0)
Irreducible error	65.58	100
	(8.11)	

model depends. For example, as we discuss in table A2, we could group subjects using auxiliary data such as their evaluations of other lotteries or response times, or use nonchoice data, such as the hypothetical choices in Bernheim et al. (2020).

We note that our completeness measure does not imply that in general CPT is a nearly complete model for predicting certainty equivalents, since the completeness measure we obtain is determined from a specific data set, so its generalizability depends on the extent to which that data are representative. However, Peysakhovich and Naecker (2017) find that CPT approximates the performance of regularized regression models for a data set of three-outcome lotteries, which suggests that our finding is robust to certain three-outcome lotteries, although the results of Bernheim and Sprenger (2020) show this will not be true for all of them.[19]

B. Across Domains

Finally, we repeat our analysis for the completeness of CPT on two additional data sets from Bruhin, Fehr-Duda, and Epper (2010). Besides the original data set, labeled "Zurich 2003" in table 5, the data sets labeled "Beijing 2005" and "Zurich 2006" respectively include 4,225 reported certainty equivalents elicited in an experiment in Beijing in 2005 (with 151 subjects) and 4,669 observations elicited in an experiment in Zurich in 2006 (with 118 subjects). The three experiments all used the same experimental design, although there was some variation in the set of lotteries.

Table 5 shows that the raw mean-squared error of CPT varies substantially across the three data sets (from 4.94 to 67.78). This contrast may suggest at face value that CPT is a more effective model of certain subject populations or lotteries than others, but the completeness of CPT turns out to be very stable across all three data sets, and is lower bounded by 92%. This comparison thus again highlights the need for appropriate benchmarks for interpreting raw prediction errors.

[19] The specification of CPT in Peysakhovich and Naecker (2017) sets $\delta = 1$ and thus has one fewer free parameter, so its model error may be higher.

TABLE 5
CPT'S ERROR VARIES SUBSTANTIALLY ACROSS DOMAINS, BUT ITS COMPLETENESS DOES NOT

	Zurich 2003	Beijing 2005	Zurich 2006
Baseline	104.63	13.23	61.43
	(10.14)	(1.41)	(6.58)
CPT	67.78	4.94	40.33
	(8.37)	(.83)	(5.86)
Irreducible error	65.58	4.89	38.36
	(7.11)	(.82)	(5.21)
Completeness (%)	94	99	92
	(2.0)	(.8)	(4.8)

C. Application 2: Initial Play in Games

1. Background and Data

In many game theory experiments, equilibrium analysis is a poor predictor of the choices that people make when they encounter a new game. This has led to models of initial play that depart from equilibrium theory, for example, the level-k models of Stahl and Wilson (1994) and Nagel (1995), the Poisson cognitive hierarchy model (Camerer, Ho, and Chong 2004), and the related models surveyed in Crawford, Costa-Gomes, and Iriberri (2013). These models represent improvements over the equilibrium predictions, but we do not know whether these models exhaust the regularities in initial play.

2. Prediction Task and Models

We consider prediction of the action chosen by the row player in a given instance of play of a 3×3 normal-form game. The available features are the 18 entries of the payoff matrix, and a prediction rule is any map $f : \mathbb{R}^{18} \to \{a_1, a_2, a_3\}$ from 3×3 payoff matrices to row player actions.

For each prediction rule f and test set of observations $\{(g_i, a_i\}_{i=1}^n$—where g_i is the payoff matrix in observation i, and a_i is the observed row player action—we evaluate error using the *misclassification rate*,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[f(g_i) \neq a_i].$$

This is the fraction of observations where the predicted action was not the observed action.

As a baseline, we consider guessing uniformly at random for all games, which yields an expected misclassification rate of 2/3. We use this benchmark to evaluate a prediction rule based on the Poisson cognitive hierarchy model (PCHM), which supposes that there is a distribution over players of differing levels of sophistication: The level-0 player randomizes

uniformly over his available actions, while the level-1 player best responds to level-0 play (Stahl and Wilson 1994, 1995; Nagel 1995). Camerer, Ho, and Chong (2004) define the play of level-k players, $k \geq 2$, to be the best response to a perceived distribution

$$p_k(h) = \frac{\pi_\tau(h)}{\sum_{l=0}^{k-1} \pi_\tau(l)} \qquad \forall\ h \in N_{<k}, \tag{8}$$

over (lower) opponent levels, where π_τ is the Poisson distribution with rate parameter τ.[20] We can derive a predicted distribution over actions by supposing that the proportion of level-k players in the population is proportional to $\pi_\tau(k)$. Assuming that this is the true distribution of play, the misclassification rate is minimized by predicting a mode of this distribution; we call this the PCHM prediction.

3. Comparison across Games

We compare the performance of the PCHM relative to the best achievable performance on three subsamples of a data set from Fudenberg and Liang (2019).[21] Our full data set consists of 23,137 total observations of initial play from 486 3×3 matrix games, where observations are pooled across all of the subjects and games.[22]

The first subsample, game set A, consists of the 16,660 observations of play from the 359 games with no strictly dominated actions.[23] Game set B consists of the 7,860 observations of play from the 161 games in which the profile that maximizes the sum of the players' payoffs is much larger (at least 20% of the largest row player payoff in the game) than the highest sum of payoffs that can be achieved when the row player chooses a level-k action (for any k).[24] For example, in the game below (which is included in game set B), the action profile (a_1, a_2) leads to a payoff sum of 160, but the

[20] Throughout, we take τ to be a free parameter and estimate it from the training data.

[21] Fudenberg and Liang (2019) studied a related prediction task, namely, predicting the modal row player action in a given game. For that prediction task, the best achievable error is always zero. Here we consider prediction of the action played, where the best achievable error depends on the true distribution of play.

[22] These data are an aggregate of three data sets: the first is a meta–data set of play in 86 games, collected from six experimental game theory papers by Kevin Leyton-Brown and James Wright (see Wright and Leyton-Brown 2014); the second is a data set of play in 200 games with randomly generated payoffs, which were gathered on Mechanical Turk for Fudenberg and Liang (2019); the third is a data set of play in 200 games that were "algorithmically designed" for a certain model (level 1) to perform poorly, again from Fudenberg and Liang (2019). There was no learning in these experiments: subjects were randomly matched to opponents, were not informed of their partners' play, and did not learn their own payoffs until the end of the session.

[23] Specifically, we consider games where no pure action is strictly dominated by another pure action.

[24] Following Stahl and Wilson (1995) and Nagel (1995), level 0 corresponds to uniform play, and each level-k action is the best response to level-$(k - 1)$ play.

largest payoff sum using level-k actions is 120. The difference, 40, is more than 20% of the maximum row player payoff in this game, 100.[25]

	a_1	a_2	a_3
a_1	40, 40	10, 20	70, 30
a_2	20, 10	80, 80	0, 100
a_3	30, 70	100, 0	60, 60

Finally, game set C consists of the 9,243 observations of play from the 175 games where the level-1 action's expected payoff against uniform play is much higher than the expected payoff of the next best action (specifically, it is larger by at least 1/4 of the maximum row player payoff in the game).

The analysis we perform for these three subsamples can be conducted for arbitrary sets of games.

4. Results

In table 6, we report the estimated irreducible error and associated completeness measures for each of the three sets of games. Our estimate for the irreducible error is derived using a lookup table, where each game is mapped to the action most commonly chosen in that game in the training data. Since we have on average 50 observations per game, the modal action in the training data is a good approximation for the modal action in the test data. High irreducible error means that there is substantial heterogeneity in play, so predicting the mode still leads to a high rate of incorrect classification. Low irreducible error means that play across subjects is more coordinated on a single action. We find that the estimated irreducible error is largest—and hence, there is the most heterogeneity in play—in data set A, which includes only games where there are no strictly dominated actions, and smallest in data set C, which includes only games where the level-1 action has by far the highest expected payoff against uniform play.

Next we use the estimated irreducible errors as a benchmark to evaluate the completeness of the PCHM on the three data sets. Although the PCHM achieves a better absolute prediction error in game set A than in game set B, its completeness is approximately 68% on both data sets. In contrast, the PCHM achieves 97% of the feasible reduction in prediction error in game set C. This means that the PCHM captures essentially all of the predictable variation in games where the level 1 action clearly has the

[25] In this game, action a_3 is level 1, since it yields the highest expected payoff against uniform play, and action a_1 is level 2, since it is the best response against play of a_3. Because (a_1, a_1) is a pure-strategy Nash equilibrium, action a_1 is level k for all $k \geq 2$. The largest payoff sum using level-k actions is achieved by the profile (a_3, a_3).

TABLE 6
COMPARISON OF THE COMPLETENESS OF THE PCHM ACROSS THE THREE SETS OF GAMES

	GAME SET A		GAME SET B		GAME SET C	
	Error	Completeness (%)	Error	Completeness (%)	Error	Completeness (%)
Baseline	.66	0	.66	0	.66	0
PCHM	.49 (.006)	68	.44 (.009)	68	.28 (.005)	97
Irreducible error	.41 (.005)	100	.34 (.006)	100	.27 (.005)	100

largest expected value against uniform play, while there is additional structure beyond the PCHM in game sets A and B. We leave to future work the question of what additional properties of the game are important determinants of the completeness of the PCHM.

D. Application 3: Human Generation of Random Sequences

1. Background and Data

Extensive experimental and empirical evidence suggests that humans misperceive randomness, for example, expecting that sequences of coin flips "self-correct" (too many heads in a row must be followed by a tails) and are balanced (the numbers of heads and tails are approximately the same) (Tversky and Kahneman 1971; Bar-Hillel and Wagenaar 1991). These misperceptions are significant not only for their basic psychological interest, but also for the ways in which misperception of randomness manifests itself in a variety of contexts: for example, investors' judgment of sequences of (random) stock returns (Barberis, Shleifer, and Vishny 1998), professional decision makers' reluctance to choose the same (correct) option multiple times in succession (Chen, Shue, and Moskowitz 2016), and people's execution of a mixed strategy in a game (Batzilis et al. 2016).

A common experimental framework in this area is to ask human participants to generate fixed-length strings of k (pseudo)random coin flips, for some small value of k (e.g., $k = 8$), and then to compare the produced distribution over length-k strings to the output of a Bernoulli process that generates realizations from $\{H, T\}$ independently and uniformly at random (Rapaport and Budescu 1997; Nickerson and Butler 2009). Following in this tradition, we use the platform Mechanical Turk to collect a large data set of human-generated strings designed to simulate the output of a Bernoulli(0.5) process, in which each symbol in the string is generated from $\{H, T\}$ independently and uniformly at random. To incentivize effort, we told subjects that payment would be approved only if their (set of) strings

could not be identified as human-generated with high confidence.[26] After removing subjects who were clearly not attempting to mimic a random process, our final data set consisted of 21,975 strings generated by 167 subjects.[27]

2. Prediction Task, Performance Metric, and Models

We consider the problem of predicting the probability that the eighth entry in a string is H given its first seven entries. To do this, we extend our framework from section III.A as follows. While the observed outcomes belong to $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{H, T\}^7$ and $\mathcal{Y} = \{H, T\}$, we consider prediction rules $f : \mathcal{X} \to [0, 1]$ that map the observed features into a probability that the final flip is H.[28]

Given a test data set $\{(s_i^1, \ldots, s_i^8)\}_{i=1}^n$ of n binary strings of length 8, we evaluate the error of the prediction rule f using the following criterion:

$$\frac{1}{n}\sum_{i=1}^{n}\left[s_i^8 - f\left(s_i^1, \ldots, s_i^7\right)\right]^2,$$

where $f(s_i^1, \ldots, s_i^7)$ is the predicted probability that the eighth flip is H given the observed initial seven flips $s_i^1, \ldots s_i^7$, and s_i^8 is the actual eighth flip. Note that the baseline of unconditionally guessing 0.5 guarantees a mean-squared prediction error of 0.25. Moreover, if the strings in the test set were truly generated via a Bernoulli(0.5) process, then no prediction rule could improve in expectation upon the baseline error.[29] We expect that behavioral errors in the generation process will make it possible to improve upon the baseline, but do not know how much it is possible to improve upon 0.25.

[26] In one experiment, 537 subjects each produced 50 binary strings of length eight. In a second experiment, an additional 101 subjects were asked to each generate 25 binary strings of length eight. Subjects were informed as follows: "To encourage effort in this task, we have developed an algorithm (based on previous Mechanical Turkers) that detects human-generated coin flips from computer-generated coin flips. You are approved for payment only if our computer is not able to identify your flips as human-generated with high confidence."

[27] Our initial data set consists of 29,375 binary strings. We chose to remove all subjects who repeated any string in more than five rounds. This cutoff was selected by looking at how often each subject generated any given string and finding the average "highest frequency" across subjects. This turned out to be 10% of the strings, or five strings. Thus, our selection criterion removes all subjects whose highest frequency was above average. This selection eliminated 167 subjects and 7,400 strings, yielding a final data set with 471 subjects and 21,975 strings. We check that our main results are not too sensitive to this selection criterion by considering two alternative choices in app. D1—first, keeping only the initial 25 strings generated by all subjects; second, removing the subjects whose strings are "most different" from a Bernoulli process under a χ^2-test. We find very similar results under these alternative criteria.

[28] As in the previous examples, we fit a representative-agent model and do not treat the identity of the subject as a feature.

[29] Due to the convexity of the loss function, it is possible to do worse than the baseline, e.g., by predicting 1 unconditionally.

In this task, the natural baseline is the rule that unconditionally guesses that the probability the final flip is H is 0.5. We compare this baseline to prediction rules based on Rabin (2002) and Rabin and Vayanos (2010), both of which predict negatively autocorrelated sequences.[30] Our prediction rule based on Rabin (2002) supposes that subjects generate sequences by drawing sequentially without replacement from an urn containing $0.5N$ "1" balls and $0.5N$ "0" balls. The urn is "refreshed" (meaning the composition is returned to its original) every period with independent probability p. This model has two free parameters: $N \in Z_+$ and $p \in [0, 1]$.

Our prediction rule based on Rabin and Vayanos (2010) assumes that the first flip $s_1 \sim \text{Bernoulli}(0.5)$ while each subsequent flip s_k is distributed

$$s_k \sim \text{Ber}\left(0.5 - \alpha \sum_{t=0}^{k-2} \delta^t (2s_{k-t-1} - 1)\right),$$

where the parameter $\delta \in \mathbb{R}_+$ reflects the (decaying) influence of past flips, and the parameter $\alpha \in \mathbb{R}_+$ measures the strength of negative autocorrelation.

3. Results

Table 7 shows that both prediction rules improve upon the baseline. The need for a benchmark for achievable prediction is starkest in this application, as the best improvement is only 0.0006, while the gap between the achieved prediction errors and a perfect zero is large. This is not surprising; since the data are generated by subjects attempting to mimic a fair coin, we naturally expect substantial variation in the eighth flip after conditioning on the initial seven flips.

For this problem, we can approximate the irreducible error by learning the empirical frequency with which each length-7 string is followed by H in the training data. Although there are 2^7 unique initial sequences, with approximately 21,000 strings in our data set we have (on average) 164 observations per initial sequence.

We find that irreducible error in this problem is 0.2441 (table 8), so that naively comparing achieved prediction error against perfect prediction (which would suggest a completeness measure for the existing models of at most 0.2%) grossly misrepresents the performance of the models. The existing models produce up to 10% of the achievable reduction in prediction error. This suggests that although negative autocorrelation is indeed present in the human-generated strings and explains a sizable

[30] Although both of these frameworks are models of mistaken inference from data, as opposed to human attempts to generate random sequences, they are easily adapted to our setting, as the papers explain.

TABLE 7
BOTH MODELS IMPROVE UPON NAIVE GUESSING, BUT THE ABSOLUTE IMPROVEMENT IS SMALL

	Error
Baseline	.25
Rabin 2002	.2496
	(.0003)
Rabin and Vayanos 2010	.2494
	(.0003)

part of the deviation from a Bernoulli(0.5) process, there is additional structure that could yet be exploited for prediction.

VI. Extensions

A. *Subject Heterogeneity*

So far, we have evaluated the completeness of "representative agent" models that implement a single prediction across all subjects. When we evaluate models that allow for subject heterogeneity, the question of what is the largest achievable reduction in prediction error is still relevant, and the irreducible error for the new expanded feature set can again help us determine the size of potential error reductions. As a simple illustration, we return to our evaluation of risk preferences and demonstrate how to construct a predictive bound for certain models with subject heterogeneity.

To do this, we calculate the completeness of the CPT specification of section V.A using the three groups of subjects identified by Bruhin, Fehr-Duda, and Epper (2010). Our approach for estimating the irreducible error is to learn the mean response for each lottery within each group, and predict those means. With sufficiently large groups, this method approximates the best possible accuracy given the identified groups.

Allowing for the parameters of CPT to vary across different subject groups weakly reduces both its completeness and the irreducible error. A priori we do not know how the sizes of these reductions compare, so the impact on completeness is ambiguous; we find that the completeness

TABLE 8
THE FEASIBLE REDUCTION IN PREDICTION ERROR OVER THE BASELINE IS SMALL IN THIS PROBLEM

	Error	Completeness (%)
Baseline	.25	0
Rabin 2002	.2496	7
	(.0003)	(3.3)
Rabin and Vayanos 2010	.2494	10
	(.0003)	(5.2)
Irreducible error	.2441	100
	(.0020)	

TABLE 9
CPT Estimated for the Three Groups of Subjects Identified in Bruhin, Fehr-Duda, and Epper (2010) Is 89% Complete

	Prediction Error	Completeness (%)
Baseline	104.63	0
	(6.5)	
CPT	60.67	89
	(5.0)	(5.8)
Irreducible error	55.42	100
	(6.2)	

of this three-group specification of CPT is comparable to the completeness of the original specification of CPT (see table 9). We note that because the same grouping assignment algorithm is used across approaches, the gap between irreducible error and the prediction errors does not shed light on how much predictions could be improved by better ways of grouping the subjects. The development of better grouping techniques is an interesting avenue for future work.[31]

B. Comparing Feature Sets

Above, we considered a fixed feature set $\mathcal{X}$, and evaluated the completeness of different models for prediction given this feature set. We can alternatively compare irreducible error across different feature sets as a way of contrasting the predictive limits of those features. We illustrate this comparison by revisiting our problem from section V.D—predicting human generation of randomness—and considering three feature sets.

The first feature set, $\mathcal{X}_{1:7}$, is our main feature set, which consists of the initial seven flips. Define $\mathcal{X}_{4:7} = \{H, T\} \times \{H, T\} \times \{H, T\}$ to be the feature set corresponding to flips 4–7, and $\mathcal{X}_H = \{0, 1, 2, \ldots, 7\}$ to be the number of H realizations in the first seven flips. Interpreted as lookup tables, these new feature sets correspond to "compressed" lookup tables built on different properties of the initial seven flips, where strings are partitioned based on certain properties. We can estimate irreducible error (table 10) by predicting the average continuation probability of H among all strings in the same partition element.

We find that the feature sets $\mathcal{X}_{4:7}$ and $\mathcal{X}_H$ achieve large fractions of the achievable improvement using the first seven flips. For example, using only the number of heads as a feature, it is possible to achieve 65% of the achievable reduction of the full structure of the initial flips. Using only the most recent three flips achieves 40% of the reduction from using all seven initial flips. On the other hand, the gap between irreducible

[31] A comparison of the irreducible error with the groups, 55.42, with the irreducible error from sec. V.A, 65.58, sheds light on the size of predictive gains achieved by Bruhin, Fehr-Duda, and Epper's (2010) method for grouping subjects.

TABLE 10
COMPARISON OF THE VALUE OF VARIOUS FEATURE SETS

	Error	Completeness (%)
Baseline	.25	0
Irreducible error for $\mathcal{X}_{4:7}$	.2478	40
	(.0004)	(4.8)
Irreducible error for $\mathcal{X}_H$	.2464	65
	(.0005)	(5.5)
Irreducible error for $\mathcal{X}_{1:7}$	.2441	100
	(.0020)	

error for $\mathcal{X}_{4:7}$ and for $\mathcal{X}_{1:7}$ demonstrates that there is predictive content in flips 1–3 beyond what is captured in flips 4–7.

The feature set $\mathcal{X}_{1:7}$ could be expanded to create richer feature sets, and it would be interesting to consider what additional features might significantly improve predictive accuracy, for example, "neuroeconomic" data such as the speed with which the strings were entered, or demographic data such as age or education.[32] The exercise in section VI.A, in which we used subject types (determined based on choices in auxiliary problems), illustrates yet another way to expand the feature set. As we have shown above, comparing irreducible error across different feature sets is one potentially useful approach for measuring the predictive value of those features.[33]

VII. Conclusion

When evaluating the predictive performance of an economic model, it is important to know not just whether the model is predictive, but also how complete its predictive performance is. Thus we should compare the prediction errors achieved by our models against the best achievable error for that problem, namely, the irreducible error. Irreducible error can be precisely estimated in certain prediction problems of interest in experimental economics. We demonstrate three settings in which completeness can help us evaluate the performance of existing models. Occasionally, as we found in section V.A, a model that has large prediction errors may nevertheless be nearly complete given its feature set.

We conclude with a brief discussion of our completeness measure, its limitations, and possibilities for extension.

Counterfactuals.—Economic models are often used to provide counterfactual predictions about the impact of new policies. Of course, if there

[32] As another example, recent work by Bernheim et al. (2020) tests how well a model of cumulative prospect theory that is trained on two-outcome lotteries predicts certainty equivalents for three-outcome lotteries. It finds that these "cross-domain" predictions can be improved using additional nonchoice features (e.g., survey responses).

[33] Note that the value of individual features will in general depend on what other features are available.

are no data about such policies, these counterfactual predictions rely on untested intuitions about the robustness of various forces that drive behavior. Suppose for example that the price variation in our data only comes from price changes by firms, and we want to predict the effect of a sales tax. We might conjecture that the price effects are the same as before, but in some cases consumers might be either more or less willing to accept a price increase imposed by the government. With or without an economic theory, any attempt to extrapolate from data in settings without sales taxes to the effects of sales taxes requires an untested hypothesis. And if we do have representative data on the past effect of sales taxes, the prediction problem does not involve a substantive counterfactual.[34]

Experimental data.—Experimental economists have a degree of control over the scope of their data that is not available in field studies. In particular, the experimentalist can choose to acquire a large number of observations for a fixed input space, so that nonparametric estimation of irreducible error for those inputs is feasible. Thus estimating completeness for laboratory data is feasible in many instances, as illustrated in the three applications in this paper. The main trade-off is between gathering more instances of observations for a given set of feature values, versus ranging over a larger set of feature values. With a sufficiently large budget, both may be possible.

Alternative measures of completeness.—In some cases, it may be possible to indirectly evaluate irreducible noise. For example, an interesting analogy to our approach to completeness is found in the literature on inheritability. Biologists have discovered a gap between two different methodologies for discovering how much of a particular outcome (say, propensity to have a disease) is heritable, dubbed the "missing heritability problem" (Manolio et al. 2009). Traditional methods of measuring heritability, such as through carefully controlled twin studies, do not attempt to isolate individual genes. Newer measurement techniques instead allow us to postulate individual genes as the carrier of heritability. Yet for many outcomes, the explanatory power of individual genes has proven far smaller (sometimes by an order of magnitude) than overall measures of heritability suggest. This gap has motivated further theorizing and measurement to isolate where the "missing heritability" may lie. Roughly speaking, the aggregate measures of heritability are in effect being used as an analog of our completeness metric for the specific gene-based theories.

Measuring portability.—One important question is how to compare the transferability of models across domains. Indeed, we may expect that economic models that are outperformed by machine-learning models in a given domain have higher transfer performance outside of the domain. In this sense, within-domain completeness may provide an insufficient

[34] Except in the trivial sense that any extrapolation from past data to future outcomes requires some form of inductive hypothesis.

measure of the "overall completeness" of the model, and we leave development of such notions to future work.

Appendix A

How Good Is Our Estimate of Irreducible Error?

In the main text, we use a lookup table to estimate irreducible error, and quantify the quality of the approximation by reporting bootstrapped standard errors. Below we supplement this with additional tests for whether the data sets we study are large enough for the out-of-sample error of the lookup table to be a good approximation for irreducible error.

In section A1, we compare the out-of-sample performance of the lookup table with that of bagged decision trees, an algorithm that works better on smaller quantities of data. We find that in each of our applications, the two prediction errors are similar, and the lookup table weakly outperforms bagged decision trees. In section A2, we study the sensitivity of the lookup table's performance to the quantity of data. The predictive accuracies achieved using our full data sets are very close to those achieved using, for example, just 70% of the data. This again suggests that only minimal improvements in predictive accuracy are feasible from further increases in data size. Finally, in section A3, we report standard errors for our completeness estimates using an approach outlined in Fudenberg, Gao, and Liang (2020).

A1. Comparison with Scalable Machine-Learning Algorithms

One way to evaluate whether the out-of-sample performance of the lookup table approximates the best possible prediction accuracy is to compare it with the performance of other machine-learning algorithms. Below we compare the lookup table with a bagged decision tree algorithm (also known as bootstrap-aggregated decision trees). This algorithm creates several bootstrapped data sets from the training data by sampling with replacement, and then trains a decision tree on each bootstrapped training set. Decision trees are nonlinear prediction models that recursively partition the feature space and learn a (best) constant prediction for each partition element. The prediction of the bagged decision tree algorithm is an aggregation of the predictions of individual decision trees. When the loss function is mean-squared error, the decision tree ensemble predicts the average of the predictions of the individual trees. When the loss function is misclassification rate, the decision tree ensemble predicts based on a majority vote across the ensemble of trees. Table A1 shows that for each prediction problem, the error of the bagged decision tree algorithm is comparable to and slightly worse than that of the lookup table.

TABLE A1
THE LOOKUP TABLE OUTPERFORMS BAGGED DECISION TREES IN EACH OF OUR PREDICTION PROBLEMS

	Risk	Games A	Games B	Games C	Sequences
Bagged decision trees	65.65	.45	.36	.29	.2442
Lookup table	65.58	.41	.34	.27	.2441

A2. Performance of the Lookup Table on Smaller Samples

We report here the lookup table's cross-validated performance on random samples of x% of our data, where $x \in \{10, 20, \ldots, 100\}$. For each x, we repeat the procedure 1,000 times, and report the average performance across iterations. We find that performance error flattens out for larger values of x, suggesting that the quantity of data we have is indeed large enough that further increases in the data size will not substantially improve predictive performance.

TABLE A2
PERFORMANCE OF LOOKUP TABLE $\hat{f}_{LT}$ USING x% OF THE DATA, AVERAGED OVER 100 ITERATIONS FOR EACH x

x (%)	Risk	Games A	Games B	Games C	Sequences
10	69.47	.4191	.3473	.2729	.2592
20	67.13	.4183	.3476	.2718	.2504
30	66.28	.4178	.3472	.2714	.2479
40	66.25	.4169	.3470	.2708	.2464
50	65.68	.4157	.3459	.2703	.2458
60	65.68	.4141	.3449	.2691	.2452
70	65.68	.4131	.3435	.2682	.2448
80	65.68	.4119	.3427	.2677	.2445
90	65.66	.4109	.3416	.2672	.2443
100	65.58	.4100	.3404	.2668	.2441

A3. Analytic Standard Errors

We report in table A3 the standard errors for our estimates of completeness, using the approach outlined in Fudenberg, Gao, and Liang (2020) (see proposition 3). These standard error estimates are slightly larger than the bootstrapped standard errors that we report in the main text.

TABLE A3
ANALYTICAL STANDARD ERRORS FOR COMPLETENESS ESTIMATES

	Completeness (%)
Application 1: Certainty equivalents:	
CPT evaluated on Zurich 2003 data	94
	(7.1)
CPT evaluated on Beijing 2005 data	99
	(7.5)
CPT evaluated on Zurich 2006 data	92
	(9.5)
Application 2: Initial play:	
PCHM evaluated on game set A	68
	(1.7)
PCHM evaluated on game set B	68
	(1.3)
PCHM evaluated on game set C	97
	(1.1)
Application 3: Random sequences:	
Rabin and Vayanos 2010	10
	(11)
Rabin 2002	7
	(13.5)

Appendix B

Relationship between Completeness and Nonparametric R^2

Suppose the loss function is mean-squared error, $\ell(y', y) = (y - y')^2$, and the baseline is the unconditional mean of the outcome variable, $f_{\text{base}}(y) = \mathbb{E}_P(y)$. Because $\mathcal{E}_P(f_{\text{base}}) = \text{var}(y)$, the R^2 for the model class is

$$R^2_\Theta = \frac{\mathcal{E}_p(f_{\text{base}}) - \mathcal{E}_p\left(f^*_\Theta\right)}{\mathcal{E}_p(f_{\text{base}})}.$$

The nonparametric R^2 is

$$R^2_{\text{nonpar}} = \frac{\mathcal{E}_p(f_{\text{base}}) - \mathcal{E}_p(f^*)}{\mathcal{E}_p(f_{\text{base}})},$$

where $f^*(x) = \mathbb{E}_P(y \mid x)$ is the conditional mean function. So our completeness measure coincides in this special case with the ratio $R^2_\Theta / R^2_{\text{nonpar}}$.[35]

Appendix C

Supplementary Material to Section V.A

C1. Alternative Baselines

We use the expected value of the lottery as a baseline in the main text. Below, we explore how completeness varies across alternative choices for the baseline.

First, we consider baselines based on three families of expected-utility models. Each of these families specifies a value function v over money, and the predicted certainty equivalent is $v^{-1}(p\, v(\bar{z}) + (1 - p)\, v(\underline{z}))$.

Power function.—The utility function over money is $v(z) = z^\alpha$ for $z \geq 0$ and $v(z) = -(-z)^\alpha$ for $z < 0$.

Constant absolute risk aversion.—The utility function over money is $v(z) = -e^{-\rho z}$ for all z.

Constant relative risk aversion.—The utility function over money is

$$v(z) = \begin{cases} \dfrac{z^{1-\gamma}}{1-\gamma} & \text{if } \gamma \neq 1, \\ \ln(z) & \text{if } \gamma = 1, \end{cases}$$

for $z \geq 0$, and $-v(-z)$ for $z < 0$.

For each baseline model, we sample 1,000 values of α, ρ, and γ from $[0, 1]$ uniformly at random, and evaluate the completeness of CPT with respect to each of these baselines. Across all of these baselines, the completeness of CPT does not fall below 0.94.

Another possibility is to allow the free parameter in the baseline models to be estimated on the training data. We report the completeness of CPT with respect to each of these baselines; again, there is very little variation.

[35] We thank an anonymous referee for making this observation.

TABLE C1
DIFFERENT CHOICES FOR THE BASELINE LEAD TO COMPARABLE LEVELS OF COMPLETENESS

	Completeness of CPT (%)
Original baseline	94
Power function	94.6
Constant absolute risk aversion	93.9
Constant relative risk aversion	94.6

Finally, as discussed in section III.C, an alternative to a user-specified baseline is the best unconditional prediction of Y, as estimated from the training data. The error of this unconditional prediction baseline (1,228) turns out to be substantially worse than the baseline that we used in the main text (103), and so CPT's completeness rises to 99.8%.

C2. Alternative Specifications of CPT

Besides the specification of CPT that we use in the main text, some other common alternatives include the original Tversky and Kahneman (1992) specification, which posits that the weighting function is

$$w(p) = \frac{p}{[p^{\gamma} + (1 - p)^{\gamma}]^{1/\gamma}},$$

and the Karmarker (1978) specification, which is equivalent to the one we use in the main text with δ set to 1. We report below the completeness of these alternative specifications of CPT.

TABLE C2
DIFFERENT SPECIFICATIONS OF CPT ALL YIELD HIGH LEVELS OF COMPLETENESS

	Completeness (%)
CPT (original)	94
	(2.0)
CPT (Karmarkar)	92
	(3.2)
CPT (Kahneman-Tversky)	71
	(2.8)

Appendix D

Supplementary Material to Section V.D

D1. Different Cuts of the Data

Initial strings only.—We repeat the analysis in section V.D using data from all subjects, but only their first 25 strings. This selection accounts for potential fatigue in

generation of the final strings, and leaves a total of 638 subjects and 15,950 strings. Prediction results for our main exercise are shown below using this alternative selection.

TABLE D1
COMPLETENESS OF THE RABIN AND VAYANOS (2010) MODEL ON THE SUBJECTS' FIRST 25 GENERATED STRINGS

	Error	Completeness (%)
Baseline	.25	0
Rabin and Vayanos 2010	.2491	5
	(.0005)	(6.8)
Irreducible error	.2326	100
	(.013)	

Removing the least random subjects.—For each subject, we conduct a χ^2-test for the null hypothesis that their strings were generated under a Bernoulli process. We order subjects by p-values and remove the 100 subjects with the lowest p-values (subjects whose generated strings were most different from what we would expect under a Bernoulli process). This leaves a total of 538 subjects and 24,550 strings. Prediction results for our main exercise are shown below using this alternative selection.

TABLE D2
COMPLETENESS OF THE RABIN AND VAYANOS (2010) MODEL ON THE DATA SET WITHOUT THE STRINGS PRODUCED BY THE 100 "LEAST RANDOM" SUBJECTS

	Error	Completeness (%)
Baseline	.25	0
Rabin and Vayanos 2010	.2492	12
	(.0003)	(4.0)
Irreducible error	.2431	100
	(.0019)	

D2. Experimental Instructions

Subjects on Mechanical Turk were presented with the following introduction screen:

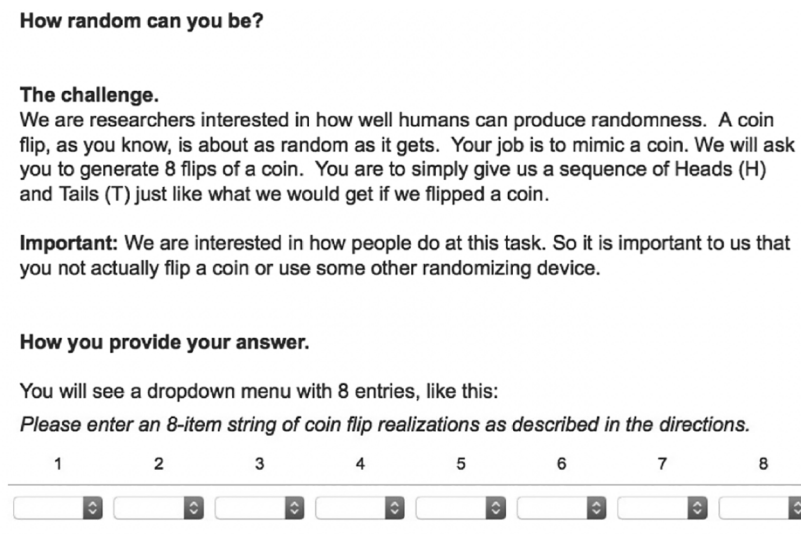

FIG. D1.—Instructions provided to subjects on Mechanical Turk.

References

Apesteguia, J., and M. Ballester. 2021. "Separating Predicted Randomness from Residual Behavior." *J. European Econ. Assoc.* 19:1041–76.

Athey, S. 2019. *The Impact of Machine Learning on Economics.* Chicago: Univ. Chicago Press.

Barberis, N., A. Shleifer, and R. Vishny. 1998. "A Model of Investor Sentiment." *J. Financial Econ.* 49:307–43.

Bar-Hillel, M., and W. Wagenaar. 1991. "The Perception of Randomness." *Advances Appl. Math.* 12:428–54.
Batzilis, D., S. Jaffe, S. Levitt, J. A. List, and J. Picel. 2016. "How Facebook Can Deepen Our Understanding of Behavior in Strategic Settings: Evidence from a Million Rock-Paper-Scissors Games." Working paper, Dept. Econ., Harvard Univ.
Bernheim, D., C. Exley, J. Naecker, and C. Sprenger. 2020. "The Model You Know: Generalizability and Predictive Power of Models of Choice under Uncertainty." Working paper, Dept. Econ., Wesleyan Univ.
Bernheim, D., and C. Sprenger. 2020. "Direct Tests of Cumulative Prospect Theory." Working paper, Dept. Econ., Harvard Univ.
Bodoh-Creed, A., J. Boenhke, and B. Hickman. 2019. "Using Machine Learning to Explain Price Dispersion." Working paper, Graduate School Management, Univ. California, Davis.
Bourgin, D. D., J. C. Peterson, D. Reichman, T. L. Griffiths, and S. J. Russell. 2019. "Cognitive Model Priors for Predicting Human Decisions." https://arxiv.org/abs/1905.09397.
Bruhin, A., H. Fehr-Duda, and T. Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica* 78:1375–1412.
Camerer, C. F., T.-H. Ho, and J.-K. Chong. 2004. "A Cognitive Hierarchy Model of Games." *Q.J.E.* 119:861–98.
Camerer, C. F., G. Nave, and A. Smith. 2019. "Dynamic Unstructured Bargaining with Private Information: Theory, Experiment, and Outcome Prediction via Machine Learning." *Management Sci.* 65:1867–90.
Chen, D., K. Shue, and T. Moskowitz. 2016. "Decision-Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *Q.J.E.* 131:1181–1242.
Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri. 2013. "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications." *J. Econ. Literature* 51:5–62.
Diebold, F., and L. Kilian. 2001. "Measuring Predictability: Theory and Macroeconomic Applications." *J. Appl. Econometrics* 16:657–69.
Erev, I., A. E. Roth, R. L. Slonim, and G. Barron. 2007. "Learning and Equilibrium as Useful Approximations: Accuracy of Prediction on Randomly Selected Constant Sum Games." *Econ. Theory* 33:29–51.
Fudenberg, D., W. Gao, and A. Liang. 2020. "Quantifying the Restrictiveness of Theories." Working paper, Dept. Econ., Massachusetts Inst. Tech.
Fudenberg, D., and A. Liang. 2019. "Predicting and Understanding Initial Play." *A.E.R.* 109:4112–41.
Gabaix, X., and D. Laibson. 2008. *The Seven Properties of Good Models.* Oxford: Oxford Univ. Press.
Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *J. American Statis. Assoc.* 102:359–78.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.
Hofman, J. M., D. J. Watts, S. Athey, et al. 2021. "Integrating Explanation and Prediction in Computational Social Science." *Nature* 595:181–88.
Karmarkar, U. 1978. "Subjectively Weighted Utility: A Descriptive Extension of the Expected Utility Model." *Org. Behavior and Human Performance* 21:67–72.
Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2018. "Human Decisions and Machine Predictions." *Q.J.E.* 133:237–93.

Manolio, T. A., F. S. Collins, N. J. Cox, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461:747–53.
Nagel, R. 1995. "Unraveling in Guessing Games: An Experimental Study." *A.E.R.* 85:1313–26.
Nickerson, R. S., and S. F. Butler. 2009. "On Producing Random Binary Sequences." *American J. Psychology* 122:141–51.
Noti, G., E. Levi, Y. Kolumbus, and A. Daniely. 2016. "Behavior-Based Machine-Learning: A Hybrid Approach for Predicting Human Decision Making." https://arxiv.org/abs/1611.10228.
Peysakhovich, A., and J. Naecker. 2017. "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity." *J. Econ. Behavior and Org.* 133:373–84.
Plonsky, O., R. Apel, E. Ert, et al. 2019. "Predicting Human Decisions with Behavioral Theories and Machine Learning." https://arxiv.org/abs/1904.06866.
Plonsky, O., I. Erev, T. Hazan, and M. Tennenholtz. 2017. "Psychological Forest: Predicting Human Behavior." *Proc. AAAI Conference Artificial Intelligence* 31:656–62.
Rabin, M. 2002. "Inference by Believers in the Law of Small Numbers." *Q.J.E.* 117:775–816.
Rabin, M., and D. Vayanos. 2010. "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Rev. Econ. Studies* 77:730–78.
Rapaport, A., and D. Budescu. 1997. "Randomization in Individual Choice Behavior." *Psychological Rev.* 104:603–17.
Samuelson, P. 1952. "Probability, Utility, and the Independence Axiom." *Econometrica* 20:670–78.
Savage, L. 1954. *The Foundations of Statistics*. New York: Wiley.
Selten, R. 1991. "Properties for a Measure of Predictive Success." *Math. Soc. Sci.* 21:153–67.
Stahl, D. O., and P. W. Wilson. 1994. "Experimental Evidence on Players' Models of Other Players." *J. Econ. Behavior and Org.* 25:309–27.
———. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Econ. Behavior* 10:218–54.
Tversky, A., and D. Kahneman. 1971. "The Belief in the Law of Small Numbers." *Psychological Bull.* 76:105–10.
———. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *J. Risk and Uncertainty* 5:297–323.
von Neumann, J., and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton Univ. Press.
Wright, J. R., and K. Leyton-Brown. 2014. "Level-0 Meta-Models for Predicting Human Behavior in Games." In *EC '14: Proceedings of the Fifteenth ACM Conference on Economics and Computation*, edited by M. Babaioff, 857–74. New York: Assoc. Computing Machinery.
Zhao, C., S. Ke, Z. Wang, and S.-L. Hsieh. 2020. "Behavioral Neural Networks." Working paper, Dept. Econ., Univ. Michigan.