Volume Title: The Economics of Artificial Intelligence: An Agenda

Volume Authors/Editors: Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-61333-8 (cloth); 978-0-226-61347-5 (electronic)

Volume URL: http://www.nber.org/books/agra-1

Conference Date: September 13–14, 2017

Publication Date: May 2019

Chapter Title: Artificial Intelligence and Behavioral Economics

Chapter Author(s): Colin F. Camerer

Chapter URL: http://www.nber.org/chapters/c14013

Chapter pages in book: (p. 587 – 608)

24

Artificial Intelligence and Behavioral Economics

Colin F. Camerer

24.1 Introduction

This chapter describes three highly speculative ideas about how artificial intelligence (AI) and behavioral economics may interact, particular in future developments in the economy and in research frontiers. First note that I will use the terms AI and machine learning (ML) interchangeably (although AI is broader) because the examples I have in mind all involve ML and prediction. A good introduction to ML for economists is Mullainathan and Spiess (2017), and other chapters in this volume.

The first idea is that ML can be used in the search for new "behavioral"-type variables that affect choice. Two examples are given, from experimental data on bargaining and on risky choice. The second idea is that some common limits on human prediction might be understood as the kinds of errors made by poor implementations of machine learning. The third idea is that it is important to study how AI technology used in firms and other institutions can both overcome and exploit human limits. The fullest understanding of this tech-human interaction will require new knowledge from behavioral economics about attention, the nature of assembled preferences, and perceived fairness.

24.2 Machine Learning to Find Behavioral Variables

Behavioral economics can be defined as the study of natural limits on computation, willpower, and self-interest, and the implications of those

Colin F. Camerer is the Robert Kirby Professor of Behavioral Finance and Economics at the California Institute of Technology.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see http://www.nber.org/chapters/c14013.ack.

limits for economic analysis (market equilibrium, IO, public finance, etc.). A different approach is to define behavioral economics more generally, as simply being open-minded about what variables are likely to influence economic choices.

This open-mindedness can be defined by listing neighboring social sciences that are likely to be the most fruitful source of explanatory variables. These include psychology, sociology (e.g., norms), anthropology (cultural variation in cognition), neuroscience, political science, and so forth. Call this the "behavioral economics trades with its neighbors" view.

But the open-mindedness could also be characterized even more generally, as an invitation to machine-learn how to predict economic outcomes from the largest possible feature set. In the "trades with its neighbors" view, features are constructs that are contributed by different neighboring sciences. These could be loss aversion, identity, moral norms, in-group preference, inattention, habit, model-free reinforcement learning, individual polygenic scores, and so forth.

But why stop there?

In a general ML approach, predictive features could be—and *should* be—*any* variables that predict. (For policy purposes, variables that could be controlled by people, firms, and governments may be of special interest.) These variables can be measurable properties of choices, the set of choices, affordances and motor interactions during choosing, measures of attention, psychophysiological measures of biological states, social influences, properties of individuals who are doing the choosing (SES, wealth, moods, personality, genes), and so forth. The more variables, the merrier.

From this perspective, we can think about what sets of features are contributed by different disciplines and theories. What features does textbook economic theory contribute? Constrained utility maximization in its most familiar and simple form points to only three kinds of variables—prices, information (which can inform utilities), and constraints.

Most propositions in behavioral economics add some variables to this list of features, such as reference-dependence, context-dependence (menu effects), anchoring, limited attention, social preference, and so forth.

Going beyond familiar theoretical constructs, the ML approach to behavioral economics specifies a very long list of candidate variables (= features) and include *all* of them in an ML approach. This approach has two advantages: First, simple theories can be seen as bets that only a small number of features will predict well; that is, some effects (such as prices) are hypothesized to be first-order in magnitude. Second, if longer lists of features predict better than a short list of theory-specified features, then that finding establishes a plausible upper bound on how much potential predictability is left to understand. The results are also likely to create raw material for theory to figure out how to consolidate the additional predictive power into crystallized theory (see also Kleinberg, Liang, and Mullainathan 2015).

If behavioral economics is recast as open-mindedness about what variables might predict, then ML is an ideal way to do behavioral economics because it can make use of a wide set of variables and select which ones predict. I will illustrate it with some examples.

Bargaining. There is a long history of bargaining experiments trying to predict what bargaining outcomes (and disagreement rates) will result from structural variables using game-theoretic methods. In the 1980s there was a sharp turn in experimental work toward noncooperative approaches in which the communication and structure of bargaining was carefully structured (e.g., Roth 1995 and Camerer 2003 for reviews). In these experiments the possible sequence of offers in the bargaining are heavily constrained and no communication is allowed (beyond the offers themselves). This shift to highly structured paradigms occurred because game theory, at the time, delivered sharp, nonobvious new predictions about what outcomes might result depending on the structural parameters—particularly, costs of delay, time horizon, the exogenous order of offers and acceptance, and available outside options (payoffs upon disagreement). Given the difficulty of measuring or controlling these structural variables in most field settings, experiments provided a natural way to test these structured-bargaining theories.[1]

Early experiments made it clear that concerns for fairness or outcomes of others influenced utility, and the planning ahead assumed in subgame perfect theories is limited and cognitively unnatural (Camerer et al. 1994; Johnson et al. 2002; Binmore et al. 2002). Experimental economists became wrapped up in understanding the nature of apparent social preferences and limited planning in structured bargaining.

However, most natural bargaining is *not* governed by rules about structure as simple as those theories, and experiments became focused from 1985 to 2000 and beyond. Natural bargaining is typically "semi-structured"—that is, there is a hard deadline and protocol for what constitutes an agreement, and otherwise there are no restrictions on which party can make what offers at what time, including the use of natural language, face-to-face meetings or use of agents, and so on.

The revival of experimental study of unstructured bargaining is a good idea for three reasons (see also Karagözoğlu, forthcoming). First, there are now a lot more ways to measure what happens during bargaining in laboratory conditions (and probably in field settings as well). Second, the large number of features that can now be generated are ideal inputs for ML to predict bargaining outcomes. Third, even when bargaining is unstructured it is possible to produce bold, nonobvious precise predictions (thanks to the revelation principle). As we will see, ML can then test whether the features

1. Examples include Binmore, Shaked, and Sutton (1985, 1989); Neelin, Sonnenschein, and Spiegel (1988); Camerer et al. (1994); and Binmore et al. (2002).

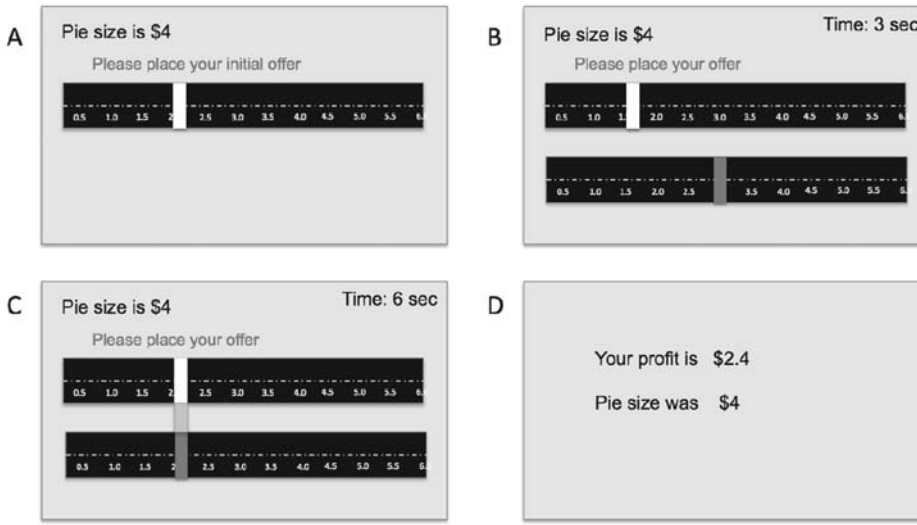

Fig. 24.1 ***A*, initial offer screen (for informed player *I*, white bar); *B*, example cursor locations after three seconds (indicating amount offered by *I*, white, or demanded by *U*, dark gray); *C*, cursor bars match which indicates an offer, consummated at six seconds; *D*, feedback screen for player *I*. Player *U* also receives feedback about pie size and profit if a trade was made (otherwise the profit is zero).**

predicted by game theory to affect outcomes actually do, and how much predictive power other features add (if any).

These three properties are illustrated by experiments of Camerer, Nave, and Smith (2017).[2] Two players bargain over how to divide an amount of money worth $1–$6 (in integer values). One informed (*I*) player knows the amount; the other, uninformed (*U*) player, doesn't know the amount. They are bargaining over how much *the uninformed U player* will get. But both players know that *I* knows the amount.

They bargain over ten seconds by moving cursors on a bargaining number line (figure 24.1). The data created in each trial is a time series of cursor locations, which are a series of step functions coming from a low offer to higher ones (representing increases in offers from *I*) and from higher demands to lower ones (representing decreasing demands from *U*).

Suppose we are trying to predict whether there will be an agreement or not based on all variables that can be observed. From a theoretical point of view, efficient bargaining based on revelation principle analysis predicts an exact rate of disagreement for each of the amounts $1–6, based only on the different amounts available. Remarkably, this prediction is process-free.

2. This paradigm builds on seminal work on semistructured bargaining by Forsythe, Kennan, and Sopher (1991).

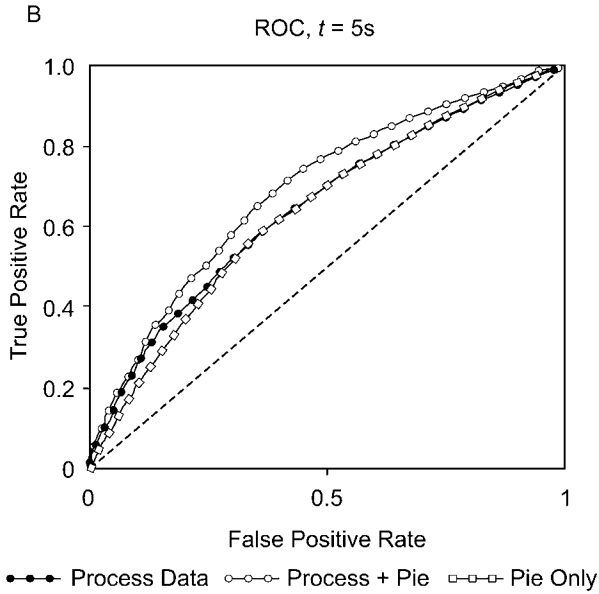

Fig. 24.2 ROC curves showing combinations of false and true positive rates in predicting bargaining disagreements

Notes: Improved forecasting is represented by curves moving to the upper left. The combination of process (cursor location features) and "pie" (amount) data are a clear improvement over either type of data alone.

However, from an ML point of view there are lots of features representing what the players are doing that could add predictive power (besides the process-free prediction based on the amount at stake). Both cursor locations are recorded every twenty-five msec. The time series of cursor locations is associated with a huge number of features—how far apart the cursors are, the time since last concession (= cursor movement), size of last concession, interactions between concession amounts and times, and so forth.

Figure 24.2 shows an ROC curve indicating test-set accuracy in predicting whether a bargaining trial ends in a disagreement (= 1) or not. The ROC curves sketch out combinations of true positive rates, *P*(disagree|predict disagree) and false positive rates *P*(agree|predict disagree). An improved ROC curve moves up and to the left, reflecting more true positives and fewer false positives. As is evident, predicting from process data only is about as accurate as using just the amount ("pie") sizes (the ROC curves with black circle and empty square markers). Using both types of data improves prediction substantially (curve with empty circle markers).

Machine learning is able to find predictive value in details of how the bargaining occurs (beyond the simple, and very good, prediction based only on the amount being bargained over). Of course, this discovery is the

beginning of the *next* step for behavioral economics. It raises questions that include: What variables predict? How do emotions,[3] face-to-face communication, and biological measures (including whole-brain imaging)[4] influence bargaining? Do people consciously understand why those variables are important? Can ML methods capture the effects of motivated cognition in unstructured bargaining, when people can self-servingly disagree about case facts?[5] Can people constrain expression of variables that hurt their bargaining power? Can mechanisms be designed that record these variables and then create efficient mediation, into which people will voluntarily participate (capturing all gains from trade)?[6]

Risky Choice. Peysakhovich and Naecker (2017) use machine learning to analyze decisions between simple financial risks. The set of risks are randomly generated triples ($y, $x, 0) with associated probabilities (p_x, p_y, p_0). Subjects give a willingness-to-pay (WTP) for each gamble.

The feature set is the five probability and amount variables (excluding the $0 payoff), quadratic terms for all five, and all two- and three-way interactions among the linear and quadratic variables. For aggregate-level estimation this creates 5 + 5 + 45 + 120 = 175 variables.

Machine learning predictions are derived from regularized regression with a linear penalty (LASSO) or squared penalty (ridge) for (absolute) coefficients. Participants were $N = 315$ MTurk subjects who each gave ten useable responses. The training set consists of 70 percent of the observations, and 30 percent are held out as a test set.

They also estimate predictive accuracy of a one-variable expected utility model (EU, with power utility) and a prospect theory (PT) model, which adds one additional parameter to allow nonlinear probability weighting (Tversky and Kahneman 1992) (with separate weights, not cumulative ones). For these models there are only one or two free parameters per person.[7]

The aggregate data estimation uses the same set of parameters for all subjects. In this analysis, the test set accuracy (mean squared error) is almost exactly the same for PT and for both LASSO and ridge ML predictions, even though PT uses only two variables and the ML methods use 175 variables. Individual-level analysis, in which each subject has their own parameters has about half the mean squared error as the aggregate analysis. The PT and ridge ML are about equally accurate.

The fact that PT and ML are equally accurate is a bit surprising because the ML method allows quite a lot of flexibility in the space of possible

3. Andrade and Ho (2009).

4. Lohrenz et al. (2007) and Bhatt et al. (2010).

5. See Babcock et al. (1995) and Babcock and Loewenstein (1997).

6. See Krajbich et al. (2008) for a related example of using neural measures to enhance efficiency in public good production experiments.

7. Note, however, that the ML feature set does not exactly nest the EU and PT forms. For example, a weighted combination of the linear outcome X and the quadratic term X^2 does not exactly equal the power function X^α.

predictions. Indeed, the authors' motivation was to use ML to show how a model with a huge amount of flexibility could fit, possibly to provide a ceiling in achievable accuracy. If the ML predictions were more accurate than EU or PT, the gap would show how much improvement could be had by more complicated combinations of outcome and probability parameters. But the result, instead, shows that much busier models are not more accurate than the time-tested two-parameter form of PT, for this domain of choices.

Limited Strategic Thinking. The concept of subgame perfection in game theory presumes that players look ahead in the future to what other players might do at future choice nodes (even choice nodes that are unlikely to be reached), in order to compute likely consequences of their current choices. This psychological presumption does have some predictive power in short, simple games. However, direct measures of attention (Camerer at al. 1994; Johnson et al. 2002) and inference from experiments (e.g., Binmore et al. 2002) make it clear that players with limited experience do not look far ahead.

More generally, in simultaneous games, there is now substantial evidence that even highly intelligent and educated subjects do not all process information in a way that leads to optimized choices given (Nash) "equilibrium" beliefs—that is, beliefs that accurately forecast what other players will do. More important, two general classes of theories have emerged that can account for deviations from optimized equilibrium theory. One class, quantal response equilibrium (QRE), are theories in which beliefs are statistically accurate but noisy (e.g., Goeree, Holt, and Palfrey 2016). Another type of theory presumes that deviations from Nash equilibrium result from a cognitive hierarchy of levels of strategic thinking. In these theories there are levels of thinking, starting from nonstrategic thinking, based presumably on salient features of strategies (or, in the absence of distinctive salience, random choice). Higher-level thinkers build up a model of what lower-level thinkers do (e.g., Stahl and Wilson 1995; Camerer, Ho, and Chong 2004; Crawford, Costa-Gomes, and Iriberri 2013). These models have been applied to hundreds of experimental games with some degree of imperfect cross-game generality, and to several field settings.[8]

Both QRE and CH/level-k theories extend equilibrium theory by adding parsimonious, precise specifications of departures from either optimization (QRE) or rationality of beliefs (CH/level-k) using a small number of behavioral parameters. The question that is asked is: Can we add predictive power in a simple, psychologically plausible[9] way using these parameters?

A more general question is: Are there structural features of payoffs and

8. For example, see Goldfarb and Xiao 2011, Östling et al. 2011, and Hortacsu et al. 2017.

9. In the case of CH/level-k theories, direct measures of visual attention from Mouselab and eyetracking have been used to test the theories using a *combination* of choices and visual attention data. See Costa-Gomes, Crawford, and Broseta 2001; Wang, Spezio, and Camerer 2010; and Brocas et al. 2014. Eyetracking and moused-based methods provide huge data

strategies that can predict even more accurately than QRE or CH/level-k? If the answer is "Yes" then the new theories, even if they are improvements, have a long way to go.

Two recent research streams have made important steps in this direction. Using methods familiar in computer science, Wright and Leyton-Brown (2014) create a "meta-model" that combines payoff features to predict what the nonstrategic "level 0" players seem to, in six sets of two-player 3 × 3 normal form games. This is a substantial improvement on previous specifications, which typically assume random behavior or some simple action based on salient information.[10]

Hartford, Wright, and Leyton-Brown (2016) go further, using deep learning neural networks (NNs) to predict human choices on the same six data sets. The NNs are able to outpredict CH models in the hold-out test sample in many cases. Importantly, even models in which there is no hierarchical iteration of strategic thinking ("layers of action response" in their approach) can fit well. This result—while preliminary—indicates that prediction purely from hidden layers of structural features can be successful.

Coming from behavioral game theory, Fudenberg and Liang (2017) explore how well ML over structural properties of strategies can predict experimental choices. They use the six data sets from Wright and Leyton-Brown (2014) and also collected data on how MTurk subjects played 200 new 3 × 3 games with randomly drawn payoffs. Their ML approach uses eighty-eight features that are categorical structural properties of strategies (e.g., Is it part of a Nash equilibrium? Is the payoff never the worst for each choice by the other player?).

The main analysis creates decision trees with k branching nodes (for k from 1 to 10) that predict whether a strategy will be played or not. Analysis uses tenfold test validation to guard against overfitting. As is common, the best-fitting trees are simple; there is a substantial improvement in fit going from $k = 1$ to $k = 2$, and then only small improvements for bushier trees. In the lab game data, the best $k = 2$ tree is simply what is called level 1 play in CH/level-k; it predicts the strategy that is a best response to uniform play by an opponent. That simple tree has a misclassification rate of 38.4 percent. The best $k = 3$ tree is only a little better (36.6 percent) and $k = 5$ is very slightly better (36.5 percent).

The model classifies rather well, but the ML feature-based models do a

sets. These previous studies heavily filter (or dimension-reduce) those data based on theory that requires consistency between choices and attention to information necessary to execute the value computation underlying the choice (Costa-Gomes, Crawford, and Broseta 2001; Costa-Gomes and Crawford 2006). Another approach that has never been tried is to use ML to select features from the huge feature set, combining choices and visual attention, to see which features predict best.

10. Examples of nonrandom behavior by nonstrategic players include bidding one's private value in an auction (Crawford and Iriberri 2007) and reporting a private state honestly in a sender-receiver game (Wang, Spezio, and Camerer 2010; Crawford 2003).

Table 24.1 **Frequency of prediction errors of various theoretical and ML models for new data from random-payoff games (from Fudenberg and Liang 2017)**

	Error	Completeness
Naïve benchmark	0.6667	1
Uniform Nash	0.4722	51.21%
	(0.0075)	
Poisson cognitive hierarchy model	0.3159	92.36%
	(0.0217)	
Prediction rule based on game features	0.2984	96.97%
	(0.0095)	
"Best possible"	0.2869	0

little better. Table 24.1 summarizes results for their new random games. The classification by Poisson cognitive hierarchy (PCH) is 92 percent of the way from random to "best possible" (using the overall distribution of actual play) in this analysis. The ML feature model is almost perfect (97 percent).

Other analyses show less impressive performance for PCH, although it can be improved substantially by adding risk aversion, and also by trying to predict different data set-specific τ values.

Note that the FL "best possible" measure is the same as the "clairvoyant" model upper bound used by Camerer, Ho, and Chong (2004). Given a data set of actual human behavior, and assuming that subjects are playing people chosen at random from that set, the best they can do is to have somehow accurately guessed what those data would be and chosen accordingly.[11] (The term "clairvoyant" is used to note that this upper bound is unlikely to be reached except by sheer lucky guessing, but if a person repeatedly chooses near the bound it implies they have an intuitive mental model of how others choose, which is quite accurate.)

Camerer, Ho, and Chong (2004) went a step further by also computing the expected *reward value* from clairvoyant prediction and comparing it with how much subjects actually earn and how much they could have earned if they obeyed different theories. Using reward value as a metric is sensible because a theory could predict frequencies rather accurately, but might not generate a much higher reward value than highly inaccurate predictions (because of the "flat maximum" property).[12] In five data sets they studied, Nash equilibrium added very little marginal value and the PCH approach

11. In psychophysics and experimental psychology, the term "ideal observer" model is used to refer to a performance benchmark closely related to what we called the clairvoyant upper bound.

12. This property was referred to as the "flat maximum" by von Winterfeldt and Edwards (1973). It came to prominence much later in experimental economics when it was noted that theories could badly predict, say, a distribution of choices in a zero-sum game, but such an inaccurate theory might not yield much less earnings than an ideal theory.

added some value in three games and more than half the maximum achievable value in two games.

24.3 Human Prediction as Imperfect Machine Learning

24.3.1 Some Pre-History of Judgment Research and Behavioral Economics

Behavioral economics as we know it and describe it nowadays, began to thrive when challenges to simple rationality principles (then called "anomalies") came to have rugged empirical status and to point to natural improvements in theory (?). It was common in those early days to distinguish anomalies about "preferences" such as mental accounting violations of fungibility and reference-dependence, and anomalies about "judgment" of likelihoods and quantities.

Somewhat hidden from economists, at that time and even now, was the fact that there was active research in many areas of judgment and decision-making (JDM). The JDM research proceeded in parallel with the emergence of behavioral economics. It was conducted almost entirely in psychology departments and some business schools, and rarely published in economics journals. The annual meeting of the S/JDM society was, for logistical efficiency, held as a satellite meeting of the Psychonomic Society (which weighted attendance toward mathematical experimental psychology).

The JDM research was about general approaches to understanding judgment processes, including "anomalies" relative to logically normative benchmarks. This research flourished because there was a healthy respect for simple mathematical models and careful testing, which enabled regularities to cumulate and gave reasons to dismiss weak results. The research community also had one foot in practical domains too (such as judgments of natural risks, medical decision-making, law, etc.) so that generalizability of lab results was always implicitly addressed.

The central ongoing debate in JDM from the 1970s on was about the cognitive processes involved in actual decisions, and the quality of those predictions. There were plenty of careful lab experiments about such phenomena, but also an earlier literature on what was then called "clinical versus statistical prediction." There lies the earliest comparison between primitive forms of ML and the important JDM piece of behavioral economics (see Lewis 2016). Many of the important contributions from this fertile period were included in the Kahneman, Slovic, and Tversky (1982) edited volume (which in the old days was called the "blue-green bible").

Paul Meehl's (1954) compact book started it all. Meehl was a remarkable character. He was a rare example, at the time, of a working clinical psychiatrist who was also interested in statistics and evidence (as were others at Minnesota). Meehl had a picture of Freud in his office, and practiced clinically for fifty years in the Veteran's Administration.

Meehl's mother had died when he was sixteen, under circumstances which apparently made him suspicious of how much doctors actually knew about how to make sick people well.

His book could be read as pursuit of such a suspicion scientifically: he collected all the studies he could find—there were twenty-two—that compared a set of clinical judgments with actual outcomes, and with simple linear models using observable predictors (some objective and some subjectively estimated).

Meehl's idea was that these statistical models could be used as a benchmark to evaluate clinicians. As Dawes and Corrigan (1974, 97) wrote, "the statistical analysis was thought to provide a floor to which the judgment of the experienced clinician could be compared. The floor turned out to be a ceiling."

In every case the statistical model outpredicted or tied the judgment accuracy of the average clinician. A later meta-analysis of 117 studies (Grove et al. 2000) found only six in which clinicians, on average, were more accurate than models (and see Dawes, Faust, and Meehl 1989).

It is possible that in any one domain, the distribution of clinicians contains some stars who could predict much more accurately. However, later studies at the individual level showed that only a minority of clinicians were more accurate than statistical models (e.g., Goldberg 1968, 1970). Kleinberg et al.'s (2017) study of machine-learned and judicial detention decisions is a modern example of the same theme.

In the decades after Meehl's book was published, evidence began to mount about *why* clinical judgment could be so imperfect. A common theme was that clinicians were good at measuring particular variables, or suggesting which objective variables to include, but were not so good at combining them consistently (e.g., Sawyer 1966). In a recollection Meehl (1986, 373) gave a succinct description of this theme:

> Why should people have been so surprised by the empirical results in my summary chapter? Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. There are no strong arguments, from the armchair or from empirical studies of cognitive psychology, for believing that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently, the query from which Lew Goldberg derived such fascinating and fundamental results.

Some other important findings emerged. One drawback of the statistical prediction approach, for practice, was that it requires large samples of high-quality outcome data (in more modern AI language, prediction required labeled data). There were rarely many such data available at the time.

Dawes (1979) proposed to give up on estimating variable weights through

a criterion-optimizing "proper" procedure like ordinary least squares (OLS),[13] using "improper" weights instead. An example is equal-weighting of standardized variables, which is often a very good approximation to OLS weighting (Einhorn and Hogarth 1975).

An interesting example of improper weights is what Dawes called "bootstrapping" (a completely distinct usage from the concept in statistics of bootstrap resampling). Dawes's idea was to regress clinical judgments on predictors, and use those estimated weights to make prediction. This is equivalent, of course, to using the predicted part of the clinical-judgment regression and discarding (or regularizing to zero, if you will) the residual. If the residual is mostly noise then correlation accuracies can be improved by this procedure, and they typically are (e.g., Camerer 1981a).

Later studies indicated a slightly more optimistic picture for the clinicians. If bootstrap-regression residuals are pure noise, they will also lower the test-retest reliability of clinical judgment (i.e., the correlation between two judgments on the same cases made by the same person). However, analysis of the few studies that report both test-retest reliability and bootstrapping regressions indicate that only about 40 percent of the residual variance is unreliable noise (Camerer 1981b). Thus, residuals do contain reliable subjective information (though it may be uncorrelated with outcomes). Blattberg and Hoch (1990) later found that for actual managerial forecasts of product sales and coupon redemption rate, residuals are correlated about .30 with outcomes. As a result, averaging statistical model forecasts and managerial judgments improved prediction substantially over statistical models alone.

24.3.2 Sparsity Is Good for You but Tastes Bad

Besides the then-startling finding that human judgment did reliably worse than statistical models, a key feature of the early results was how well small numbers of variables could fit. Some of this conclusion was constrained by the fact that there were not huge feature sets with truly large number of variables in any case (so you couldn't possibly know, at that time, if "large numbers of variables fit surprisingly better" than small numbers).

A striking example in Dawes (1979) is a two-variable model predicting marital happiness: the rate of lovemaking minus the rate of fighting. He reports correlations of .40 and .81 in two studies (Edwards and Edwards 1977; Thornton 1977).[14]

In another more famous example, Dawes (1971) did a study about admitting students to the University of Oregon PhD program in psychology from 1964 to 1967. He compared and measured each applicant's GRE, undergraduate GPA, and the quality of the applicant's undergraduate school. The

13. Presciently, Dawes also mentions using ridge regression as a proper procedure to maximize out-of-sample fit.

14. More recent analyses using transcribed verbal interactions generate correlations for divorce and marital satisfaction around .6–.7. The core variables are called the "four horsemen" of criticism, defensiveness, contempt, and "stonewalling" (listener withdrawal).

variables were standardized, then weighted equally. The outcome variable was faculty ratings in 1969 of how well the students they had admitted succeeded. (Obviously, the selection effect here makes the entire analysis much less than ideal, but tracking down rejected applicants and measuring their success by 1969 was basically impossible at the time.)

The simple three-variable statistical model correlated with later success in the program more highly (.48, cross-validated) than the admissions committee's quantitative recommendation (.19).[15] The bootstrapping model of the admissions committee correlated .25.

Despite Dawes's evidence, I have never been able to convince any graduate admissions committee at two institutions (Penn and Caltech) to actually compute statistical ratings, even as a way to filter out applications that are likely to be certain rejections.

Why not?

I think the answer is that the human mind rebels against regularization and the resulting sparsity. We are born to overfit. Every AI researcher knows that including fewer variables (e.g., by giving many of them zero weights in LASSO, or limiting tree depth in random forests) is a useful all-purpose prophylactic for overfitting a training set. But the same process seems to be unappealing in our everyday judgment.

The distaste for sparsity is ironic because, in fact, the brain is built to do a massive amount of filtering of sensory information (and does so remarkably efficiently in areas where optimal efficiency can be quantified, such as vision; see Doi et al. [2012]). But people do not like to *explicitly* throw away information (Einhorn 1986). This is particularly true if the information is already in front of us—in the form of a PhD admissions application, or a person talking about their research in an AEA interview hotel room. It takes some combination of willpower, arrogance, or what have you, to simply ignore letters of recommendation, for example. Another force is "illusory correlation," in which strong prior beliefs about an association bias encoding or memory so that the prior is maintained, incorrectly (Chapman and Chapman 1969; Klayman and Ha 1985).

The poster child for misguided sparsity rebellion is personal short face-to-face interviews in hiring. There is a mountain of evidence that such interviews do not predict anything about later work performance, if interviewers are untrained and do not use a structured interview format, that isn't better predicted by numbers (e.g., Dana, Dawes, and Peterson 2013).

A likely example is interviewing faculty candidates with new PhDs in hotel suites at the ASSA meetings. Suppose the goal of such interviews is to predict which new PhDs will do enough terrific research, good teaching,

15. Readers might guess that the quality of econometrics for inference in some of these earlier papers is limited. For example, Dawes (1971) only used the 111 students who had been admitted to the program and stayed enrolled, so there is likely scale compression and so forth. Some of the faculty members rating those students were probably also initial raters, which could generate consistency biases, and so forth.

and other kinds of service and public value to get tenure several years later at the interviewers' home institution.

That predictive goal is admirable, but the brain of an untrained interviewer has more basic things on its mind. Is this person well dressed? Can they protect me if there is danger? Are they friend or foe? Does their accent and word choice sound like mine? Why are they stifling a yawn?—they'll *never* get papers accepted at *Econometrica* if they yawn after a long tense day slipping on ice in Philadelphia rushing to avoid being late to a hotel suite!

People who do these interviews (including me) **say** that we are trying to probe the candidate's depth of understanding about their topic, how promising their new planned research is, and so forth. But what we really are evaluating is probably more like "Do they belong in my tribe?"

While I do think such interviews are a waste of time,[16] it is *conceivable* that they generate valid information. The problem is that interviewers may weight the wrong information (as well as overweighting features that should be regularized to zero). If there is valid information about long-run tenure prospects and collegiality, the best method to capture such information is to videotape the interview, combine it with other tasks that more closely resemble work performance (e.g., have them review a difficult paper), and machine learn the heck out of that larger corpus of information.

Another simple example of where ignoring information is counterintuitive is captured by the two modes of forecasting that Kahneman and Lovallo (1993) wrote about. They called the two modes the "inside" and "outside" view. The two views were in the context of forecasting the outcome of a project (such as writing a book, or a business investment). The inside view "focused only on a particular case, by considering the plan and its obstacles to completion, by constructing scenarios of future progress" (25). The outside view "focuses on the statistics of a class of cases chosen to be similar in relevant respects to the current one" (25).

The outside view deliberately throws away most of the information about a specific case at hand (but keeps some information): it reduces the relevant dimensions to *only* those that are present in the outside view reference class. (This is, again, a regularization that zeros out all the features that are not "similar in relevant respects.")

In ML terms, the outside and inside views are like different kinds of cluster analyses. The outside view parses all previous cases into K clusters; a current case belongs to one cluster or another (though there is, of course, a degree of cluster membership depending on the distance from cluster centroids). The inside view—in its extreme form—treats each case, like fingerprints and snowflakes, as unique.

16. There are many caveats, of course, to this strong claim. For example, often the school is pitching to attract a highly desirable candidate, not the other way around.

24.3.3 Hypothesis: Human Judgment Is Like Overfitted Machine Learning

The core idea I want to explore is that some aspects of everyday human judgment can be understood as the type of errors that would result from badly done machine learning.[17] I will focus on two aspects: overconfidence and how it increases, and limited error correction.

In both cases, I have in mind a research program that takes data on human predictions and compares them with machine-learned predictions. Then *deliberately* re-do the machine learning badly (e.g., failing to correct for overfitting) and see whether the impaired ML predictions have some of the properties of human ones.

Overconfidence. In a classic study from the early days of JDM, Oskamp (1965) had eight experienced clinical psychologists and twenty-four graduate and undergraduate students read material about an actual person, in four stages. The first stage was just three sentences giving basic demographics, education, and occupation. The next three stages were one and a half to two pages each about childhood, schooling, and the subject's time in the army and beyond. There were a total of five pages of material.

The subjects had to answer twenty-five personality questions about the subject, each with five multiple-choice answers[18] after each of the four stages of reading. All these questions had correct answers, based on other evidence about the case. Chance guessing would be 20 percent accurate.

Oskamp learned two things: First, there was no difference in accuracy between the experienced clinicians and the students.

Second, all the subjects were barely above chance, and accuracy did not improve as they read more material in the three stages. After just the first paragraph, their accuracy was 26 percent; after reading all five additional pages across the three stages, accuracy was 28 percent (an insignificant difference from 26 percent). However, the subjects' subjective *confidence* in their accuracy rose almost linearly as they read more, from 33 percent to 53 percent.[19]

This increase in confidence, combined with no increase in accuracy, is reminiscent of the difference between training set and test set accuracy in AI. As more and more variables are included in a training set, the (unpenalized) accuracy will always increase. As a result of overfitting, however, test-set accuracy will *decline* when too many variables are included. The

17. My intuition about this was aided by Jesse Shapiro, who asked a well-crafted question pointing straight in this direction.

18. One of the multiple choice questions was "Kid's present attitude toward his mother is one of: (a) love and respect for her ideals, (b) affectionate tolerance for her foibles," and so forth.

19. Some other results comparing more and less experienced clinicians, however, have also confirmed the first finding (experience does not improve accuracy much), but found that experience tends to *reduce* overconfidence (Goldberg 1959).

resulting gap between training- and test-set accuracy will grow, much as the overconfidence in Oskamp's subjects grew with the equivalent of more "variables" (i.e., more material on the single person they were judging).

Overconfidence comes in different flavors. In the predictive context, we will define it as having too narrow a confidence interval around a prediction. (In regression, for example, this means underestimating the standard error of a conditional prediction $P(Y|X)$ based on observables X.)

My hypothesis is that human overconfidence results from a failure to winnow the set of predictors (as in LASSO penalties for feature weights). Overconfidence of this type is a consequence of not anticipating overfitting. High training-set accuracy corresponds to confidence about predictions. Overconfidence is a failure to anticipate the drop in accuracy from training to test.

Limited Error Correction. In some ML procedures, training takes place over trials. For example, the earliest neural networks were trained by making output predictions based on a set of node weights, then back-propagating prediction errors to adjust the weights. Early contributions intended for this process to correspond to human learning—for example, how children learn to recognize categories of natural objects or to learn properties of language (e.g., Rumelhart and McClelland 1986).

One can then ask whether some aspects of adult human judgment correspond to poor implementation of error correction. An invisible assumption that is, of course, part of neural network training is that output errors are recognized (if learning is supervised by labeled data). But what if humans do not recognize error or respond to it inappropriately?

One maladaptive response to prediction error is to add features, particularly interaction effects. For example, suppose a college admissions director has a predictive model and thinks students who play musical instruments have good study habits and will succeed in the college. Now a student comes along who plays drums in the Dead Milkmen punk band. The student gets admitted (because playing music is a good feature), but struggles in college and drops out.

The admissions director could back-propagate the predictive error to adjust the weights on the "plays music" feature. Or she could create a new feature by splitting "plays music" into "plays drums" and "plays nondrums" and ignore the error. This procedure will generate too many features and will not use error-correction effectively.[20]

Furthermore, note that a *different* admissions director might create two different subfeatures, "plays music in a punk band" and "plays nonpunk music." In the stylized version of this description, both will become convinced that they have improved their mental models and will retain high confidence about future predictions. But their inter-rater reliability will have

20. Another way to model this is as the refinement of a prediction tree, where branches are added for new feature when predictions are incorrect. This will generate a bushy tree, which generally harms test-set accuracy.

gone *down*, because they "improved" their models in different ways. Inter-rate reliability puts a hard upper bound on how good average predictive accuracy can be. Finally, note that even if human experts are mediocre at feature selection or create too many interaction effects (which ML regularizes away), they are often more rapid than novices (for a remarkable study of actual admissions decisions, see Johnson 1980, 1988). The *process* they use is rapid, but the predictive *performance* is not so impressive. But AI algorithms are even faster.

24.4 AI Technology as a Bionic Patch, or Malware, for Human Limits

We spend a lot of time in behavioral economics thinking about how political and economic systems either exploit bad choices or help people make good choices. What behavioral economics has to offer to this general discussion is to specify a more psychologically accurate model of human choice and human nature than the caricature of constrained utility-maximization (as useful as it has been).

Artificial intelligence enters by creating better tools for making inferences about what a person wants and what a person will do. Sometimes these tools will hurt and sometimes they will help.

Artificial Intelligence Helps. A clear example is recommender systems. Recommender systems use previous data on a target person's choices and ex post quality ratings, as well as data on many other people, possible choices, and ratings, to predict how well the target person will like a choice they have not made before (and may not even know exists, such as movies or books they haven't heard of). Recommender systems are a behavioral prosthetic to remedy human limits on attention and memory and the resulting incompleteness of preferences.

Consider Netflix movie recommendations. Netflix uses a person's viewing and ratings history, as well as opinions of others and movie properties, as inputs to a variety of algorithms to suggest what content to watch. As their data scientists explained (Gomez-Uribe and Hunt 2016):

> a typical Netflix member loses interest after perhaps 60 to 90 seconds of choosing, having reviewed 10 to 20 titles (perhaps 3 in detail) on one or two screens. . . . The recommender problem is to make sure that on those two screens each member in our diverse pool will find something compelling to view, and will understand why it might be of interest.

For example, their "Because You Watched" recommender line uses a video-video similarity algorithm to suggest unwatched videos similar to ones the user watched and liked.

There are so many interesting implications of these kinds of recommender systems for economics in general, and for behavioral economics in particular. For example, Netflix wants its members to "understand *why* it (a recommended video) might be of interest." This is, at bottom, a ques-

tion about interpretability of AI output, how a member learns from recommender successes and errors, and whether a member then "trusts" Netflix in general. All these are psychological processes that may also depend heavily on design and experience features (UD, UX).

Artificial Intelligence "Hurts."[21] Another feature of AI-driven personalization is price discrimination. If people do know a lot about what they want, and have precise willingness-to-pay (WTP), then companies will quickly develop the capacity to personalize prices too. This seems to be a concept that is emerging rapidly and desperately needs to be studied by industrial economists who can figure out the welfare implications.

Behavioral economics can play a role by using evidence about how people make judgments about fairness of prices (e.g., Kahneman, Knetsch, and Thaler 1986), whether fairness norms adapt to "personalized pricing," and how fairness judgments influence behavior.

My intuition (echoing Kahneman, Knetsch, and Thaler 1986) is that people can come to accept a high degree of variation in prices for what is essentially the same product as long as there is either (a) very minor product differentiation[22] or (b) firms can articulate why different prices are fair. For example, price discrimination might be framed as cross-subsidy to help those who can't afford high prices.

It is also likely that personalized pricing will harm consumers who are the most habitual or who do not shop cleverly, but will help savvy consumers who can hijack the personalization algorithms to look like low WTP consumers and save money. See Gabaix and Laibson (2006) for a carefully worked-out model about hidden ("shrouded") product attributes.

24.5 Conclusion

This chapter discussed three ways in which AI, particularly machine learning, connect with behavioral economics. One way is that ML can be used to mine the large set of features that behavioral economists think *could* improve prediction of choice. I gave examples of simple kinds of ML (with much smaller data sets than often used) in predicting bargaining outcomes, risky choice, and behavior in games.

The second way is by construing typical patterns in human judgment as the output of implicit machine-learning methods that are inappropriately applied. For example, if there is no correction for overfitting, then the gap

21. I put the word "hurts" in quotes here as a way to conjecture, through punctuation, that in many industries the AI-driven capacity to personalize pricing will harm consumer welfare overall.

22. A feature of their fairness framework is that people do not mind price increases or surcharges if they are even partially justified by cost differentials. I have a recollection of Kahneman and Thaler joking that a restaurant could successfully charge higher prices on Saturday nights if there is some enhancement, such as a mariachi band—even if most people don't like mariachi.

between training set accuracy and test-set accuracy will grow and grow if more features are used. This could be a model of human overconfidence.

The third way is that AI methods can help people "assemble" preference predictions about unfamiliar products (e.g., through recommender systems) and can also harm consumers by extracting more surplus than ever before (through better types of price discrimination).

References

Andrade, E. B., and T.-H. Ho. 2009. "Gaming Emotions in Social Interactions." *Journal of Consumer Research* 36 (4): 539–52.

Babcock, L., and G. Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives* 11 (1): 109–26.

Babcock, L., G. Loewenstein, S. Issacharoff, and C. Camerer. 1995. "Biased Judgments of Fairness in Bargaining." *American Economic Review* 85 (5): 1337–43.

Bhatt, M. A., T. Lohrenz, C. F. Camerer, and P. R. Montegue. 2010. "Neural Signatures of Strategic Types in a Two-Person Bargaining Game." *Proceedings of the National Academy of Sciences* 107 (46): 19720–25.

Binmore, K., J. McCarthy, G. Ponti, A. Shaked, and L. Samuelson. 2002. "A Backward Induction Experiment." *Journal of Economic Theory* 184:48–88.

Binmore, K., A. Shaked, and J. Sutton. 1985. "Testing Noncooperative Bargaining Theory: A Preliminary Study." *American Economic Review* 75 (5): 1178–80.

———. 1989. "An Outside Option Experiment." *Quarterly Journal of Economics* 104 (4): 753–70.

Blattberg, R. C., and S. J. Hoch. 1990. "Database Models and Managerial Intuition: 50% Database + 50% Manager." *Management Science* 36 (8): 887–99.

Brocas, Isabelle, J. D. Carrillo, S. W. Wang, and C. F. Camerer. 2014. "Imperfect Choice or Imperfect Attention? Understanding Strategic Thinking in Private Information Games." *Review of Economic Studies* 81 (3): 944–70.

Camerer, C. F. 1981a. "General Conditions for the Success of Bootstrapping Models." *Organizational Behavior and Human Performance* 27:411–22.

———. 1981b. "The Validity and Utility of Expert Judgment." Unpublished PhD diss., Center for Decision Research, University of Chicago Graduate School of Business.

———. 2003. *Behavioral Game Theory, Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Camerer, C. F., T.-H. Ho, and J.-K. Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119 (3): 861–98.

Camerer, C., E. Johnson, T. Rymon, and S. Sen. 1994. "Cognition and Framing in Sequential Bargaining for Gains and Losses. In *Frontiers of Game Theory*, edited by A. Kirman, K. Binmore, and P. Tani, 101–20. Cambridge, MA: MIT Press.

Camerer, C. F., G. Nave, and A. Smith. 2017. "Dynamic Unstructured Bargaining with Private Information and Deadlines: Theory and Experiment." Working paper.

Chapman, L. J., and J. P. Chapman. 1969. "Illusory Correlation as an Obstacle to the Use of Valid Psychodiagnostic Signs." *Journal of Abnormal Psychology* 46:271–80.

Costa-Gomes, M. A., and V. P. Crawford. 2006. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." *American Economic Review* 96 (5): 1737–68.

Costa-Gomes, M. A., V. P. Crawford, and B. Broseta. 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica* 69 (5): 1193–235.

Crawford, V. P. 2003. "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions." *American Economic Review* 93 (1): 133–49.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri. 2013. "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications." *Journal of Economic Literature* 51 (1): 5–62.

Crawford, V. P., and N. Iriberri. 2007. "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica* 75 (6): 1721–70.

Dana, J., R. Dawes, and N. Peterson. 2013. "Belief in the Unstructured Interview: The Persistence of an Illusion." *Judgment and Decision Making* 8 (5): 512–20.

Dawes, R. M. 1971. "A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making." *American Psychologist* 26:180–88.

———. 1979. "The Robust Beauty of Improper Linear Models in Decision Making." *American Psychologist* 34 (7): 571.

Dawes, R. M., and B. Corrigan. 1974. "Linear Models in Decision Making." *Psychological Bulletin* 81, 97.

Dawes, R. M., D. Faust, and P. E. Meehl. 1989. "Clinical versus Actuarial Judgment." *Science* 243:1668–74.

Doi, E., J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, T. A. Machado, et al. 2012. "Efficient Coding of Spatial Information in the Primate Retina." *Journal of Neuroscience* 32 (46): 16256–64.

Edwards, D. D., and J. S. Edwards. 1977. "Marriage: Direct and Continuous Measurement." *Bulletin of the Psychonomic Society* 10:187–88.

Einhorn, H. J. 1986. "Accepting Error to Make Less Error." *Journal of Personality Assessment* 50:387–95.

Einhorn, H. J., and R. M. Hogarth. 1975. "Unit Weighting Schemas for Decision Making." *Organization Behavior and Human Performance* 13:171–92.

Forsythe, R., J. Kennan, and B. Sopher. 1991. "An Experimental Analysis of Strikes in Bargaining Games with One-Sided Private Information." *American Economic Review* 81 (1): 253–78.

Fudenberg, D., and A. Liang. 2017. "Predicting and Understanding Initial Play." Working paper, Massachusetts Institute of Technology and the University of Pennsylvania.

Gabaix, X., and D. Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics* 121 (2): 505–40.

Goeree, J., C. Holt, and T. Palfrey. 2016. *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton, NJ: Princeton University Press.

Goldberg, L. R. 1959. "The Effectiveness of Clinicians' Judgments: The Diagnosis of Organic Brain Damage from the Bender-Gestalt Test." *Journal of Consulting Psychology* 23:25–33.

———. 1968. "Simple Models or Simple Processes?" *American Psychologist* 23:483–96.

———.1970. "Man versus Model of Man: A Rationale, Plus Some Evidence for a Method of Improving on Clinical Inferences." *Psychological Bulletin* 73:422–32.

Goldfarb, A., and M. Xiao. 2011. "Who Thinks about the Competition? Managerial Ability and Strategic Entry in US Local Telephone Markets." *American Economic Review* 101 (7): 3130–61.

Gomez-Uribe, C., and N. Hunt. 2016. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." *ACM Transactions on Management Information Systems (TMIS)* 6 (4): article 13.

Grove, W. M., D. H. Zald, B. S. Lebow, B. E. Snits, and C. E. Nelson. 2000. "Clinical vs. Mechanical Prediction: A Meta-analysis." *Psychological Assessment* 12:19–30.

Hartford, J. S., J. R. Wright, and K. Leyton-Brown. 2016. "Deep Learning for Pre-

dicting Human Strategic Behavior." *Advances in Neural Information Processing Systems*. https://dl.acm.org/citation.cfm?id=3157368.
Hortacsu, A., F. Luco, S. L. Puller, and D. Zhu. 2017. "Does Strategic Ability Affect Efficiency? Evidence from Electricity Markets." NBER Working Paper no. 23526, Cambridge, MA.
Johnson, E. J. 1980. "Expertise in Admissions Judgment." Unpublished PhD diss., Carnegie-Mellon University.
Johnson, E. J. 1988. "Expertise and Decision under Uncertainty: Performance and Process." In *The Nature of Expertise*, edited by M. T. H. Chi, R. Glaser, and M. I. Farr, 209–28. Hillsdale, NJ: Erlbaum.
Johnson, E. J., C. F. Camerer, S. Sen, and T. Rymon. 2002. "Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining." *Journal of Economic Theory* 104 (1): 16–47.
Kahneman, D., J. L. Knetsch, and R. Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* : 728–41.
Kahneman, D., and D. Lovallo. 1993. "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking." *Management Science* 39 (1): 17–31.
Kahneman, D., P. Slovic, and A. Tversky, eds. 1982. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.
Karagözoğlu, E. Forthcoming. "On 'Going Unstructured' in Bargaining Experiments." *Studies in Economic Design by Springer, Future of Economic Design.*
Klayman, J., and Y. Ha. 1985. "Confirmation, Disconfirmation, and Information in Hypothesis Testing." *Psychological Review* : 211–28.
Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2017. "Human Decisions and Machine Predictions." NBER Working Paper no. 23180, Cambridge, MA.
Kleinberg, J., A. Liang, and S. Mullainathan. 2015. "The Theory is Predictive, But Is It Complete? An Application to Human Perception of Randomness." Unpublished manuscript..
Krajbich, I., C. Camerer, J. Ledyard, and A. Rangel. 2009. "Using Neural Measures of Economic Value to Solve the Public Goods Free-Rider Problem." *Science* 326 (5952): 596–99.
Lewis, M. 2016. *The Undoing Project: A Friendship That Changed Our Minds.* New York: W. W. Norton.
Lohrenz, T., J. McCabe, C. F. Camerer, and P. R. Montague. 2007. "Neural Signature of Fictive Learning Signals in a Sequential Investment Task." *Proceedings of the National Academy of Sciences* 104 (22): 9493–98.
Meehl, P. E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis: University of Minnesota Press.
———. 1986. "Causes and Effects of My Disturbing Little Book." *Journal of Personality Assessment* 50 (3): 370–75.
Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.
Neelin, J., H. Sonnenschein, and M. Spiegel. 1988. "A Further Test of Noncooperative Bargaining Theory: Comment." *American Economic Review* 78 (4): 824–36.
Oskamp, S. 1965. "Overconfidence in Case-Study Judgments." *Journal of Consulting Psychology* 29 (3): 261.
Östling, R., J. Wang, E. Chou, and C. F. Camerer. 2011. "Strategic Thinking and Learning in the Field and Lab: Evidence from Poisson LUPI Lottery Games." *American Economic Journal: Microeconomics* 23 (3): 1–33.
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* 133:373–84.
Roth, A. E. 1995. "Bargaining Experiments." In *Handbook of Experimental Econom-*

ics, edited by J. Kagel and A. Roth, 253–348. Princeton, NJ: Princeton University Press.
Rumelhart, D. E., and J. L. McClelland. 1986. "On Learning the Past Tenses of English Verbs." In *Parallel Distributed Processing*, vol. 2, edited by D. Rumelhart, J. McClelland, and the PDP Research Group, 216–71. Cambridge, MA: MIT Press.
Sawyer, J. 1966. "Measurement and Prediction, Clinical and Statistical." *Psychological Bulletin* 66:178–200.
Stahl, D. O., and P. W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10 (1): 218–54.
Thornton, B. 1977. "Linear Prediction of Marital Happiness: A Replication." *Personality and Social Psychology Bulletin* 3:674–76.
Tversky, A., and D. Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5 (4): 297–323.
von Winterfeldt, D., and W. Edwards. 1973. "Flat Maxima in Linear Optimization Models." Working Paper no. 011313-4-T, Engineering Psychology Lab, University of Michigan, Ann Arbor.
Wang, J., M. Spezio, and C. F. Camerer. 2010. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games." *American Economic Review* 100 (3): 984–1007.
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*, 857–74.

Comment Daniel Kahneman

Below is a slightly edited version of Professor Kahneman's spoken remarks.

During the talks yesterday, I couldn't understand most of what was going on, and yet I had the feeling that I was learning a lot. I will have some remarks about Colin (Camerer) and then some remarks about the few things that I noticed yesterday that I could understand.

Colin had a lovely idea that I agree with. It is that if you have a mass of data and you use deep learning, you will find out much more than your theory is designed to explain. And I would hope that machine learning can be a source of hypotheses. That is, that some of these variables that you identify are genuinely interesting.

At least in my field, the bar for successful publishable science is very low. We consider theories confirmed even when they explain very little of the variance so long as they yield statistically significant predictions. We treat the residual variance as noise, so a deeper look into the residual variance, which machine learning is good at, is an advantage. So as an outsider, actu-

Daniel Kahneman is professor emeritus of psychology and public affairs at the Woodrow Wilson School and the Eugene Higgins Professor of Psychology emeritus, Princeton University, and a fellow of the Center for Rationality at the Hebrew University in Jerusalem.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see http://www.nber.org/chapters/c14016.ack.