Artificial intelligence and behavioral economics

Colin F. Camerer
Caltech
camerer@caltech.edu

9/7/17. 10:01am PST. Prepared for NBER conference, September 13-14, 2017.
 Comments are very much welcome.

I: Introduction

This paper describes 2-1/2 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'll 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).

The first idea is that AI 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. That is, people are thinking as if they are executing machine learning algorithms but are doing a mediocre job of it.

The half idea— it's short-- is that it is important to study how AI technology used in firms and by 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 and perceived fairness.

Please accept my apologies for limits of this paper: I know a bit about AI, from teaching neural networks in the 1990s, teaching data analytics more recently, and having generous smart colleagues at Caltech and elsewhere, as teachers. But I'm not an expert. And I wrote this version in a hurry!

II: 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 limits for market equilibrium. A different approach is to define behavioral economics more

generally, as simply being open-minded about what variables are likely to influence economic choices.

One way to describe this open-mindedness is to list neighboring social sciences which are likely to be the most fruitful source of explanatory variables—psychology, perhaps sociology (e.g., norms), anthropology (cultural variation in cognition), neuroscience, etc. Call this the "behavioral economics borrows from its neighbors" view.

But the open-mindedness could also be characterized even more generally, as an invitation to machine-learn how to predict from the largest possible feature set. In the "behavioral economics borrows from its neighbors" view, features are constructs and their measures contributed by different sciences. These could be loss-aversion, identity, moral norms, in-group preference, inattention, habit, model-free reinforcement learning, etc.

But why stop there?

In a general ML approach, predictive features could be—and *should* be-- *any* variables that predict. These 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 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 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, etc.

Another way to search for predictive variables is to specify a very long list of candidate variables (=features) and include *all* of them in an ML approach[1]. This approach has two advantages: First, simple theories can be seen as bets that only a small number of features will predict well. 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).

[1] Another example is using multivoxel pattern analysis to "decode" brain activity for

If behavioral economics is thought of as open-mindedness about what variables might predict (as it is for me) then an ML approach with many variables is a potentially useful approach. I'll illustrate it with some examples.[2]

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, towards noncooperative approaches in which the communication and structure of bargaining was carefully fixed. This happened because game theory, at the time, delivered sharp new predictions about what outcomes might result depending on the structural parameters (such as the order of bargaining offers, discount rates over time, etc.)

However, most natural bargaining is not governed by such strict rules. Therefore, it is important to understand what bargaining outcomes might result when bargaining is "semi-structured". "Semi-structured" means there is a deadline and protocol for acceptance but otherwise no restrictions on who can offer what at what time (and potentially including language).

Unstructured bargaining is ripe for machine learning. In the experiments of Camerer et al (2017), 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, uniformed (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 10 second by moving cursors on a bargaining number line (Figure 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).

[2] Fudenberg-Liang add later

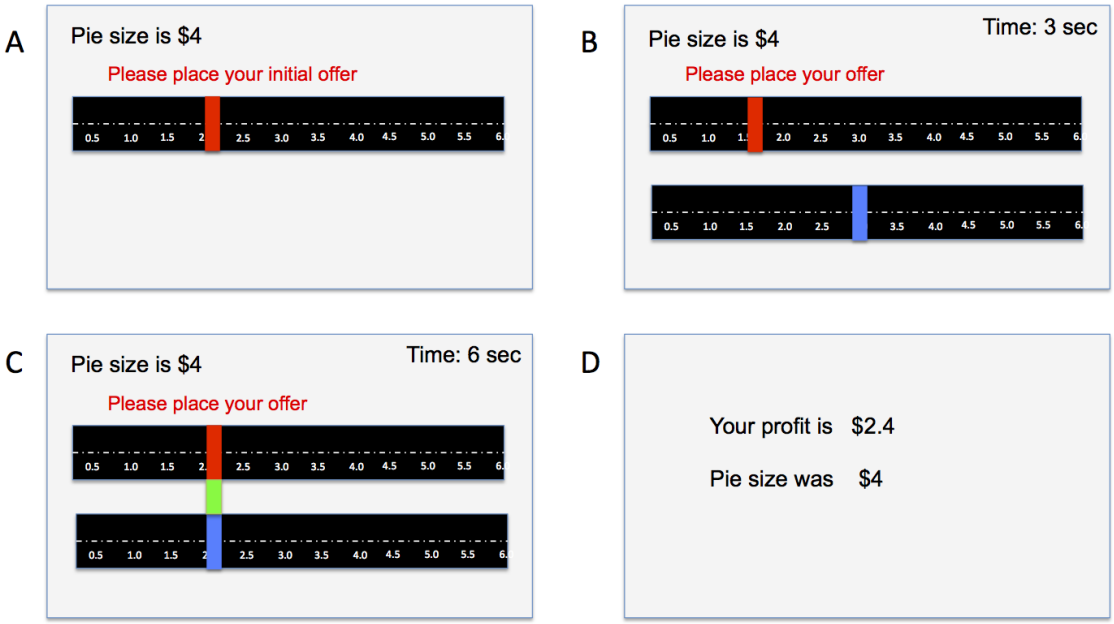

Figure 1: (A) Initial offer screen (for informed player I, red bar); (B) example cursor locations after 3 secs (indicating amount offered by I, red, or demanded by U, blue). (C) cursor bars match which indicates an offer, consummated at 6 seconds. (D) Feedback screen for player I. Player U also receives feedback about pie size and profit if a trade was made (otherwise zero).

Suppose are trying to predict whether there will be an agreement or not based on everything 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.

However, from an ML point of view there are lots of features representing what the players are doing which could add predictive power (besides the process-free prediction based on the amount at stake). Both cursor locations are recorded every 25 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, etc.

Ongoing experiments also record facial expressions and psychophysiology (skin conductance). The main point to appreciate is that from an ML point of view, recording everything possible during the bargaining, and about the players (personality, wealth, mood) could all potentially improve prediction of whether there is disagreement.

Figure 2 shows an ROC curve indicating test-set accuracy in predicting whether a bargaining trial ends in a disagreement (=1) or not. 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 blue and green ROC curves). Using both types of data improves prediction substantially (the red curve).

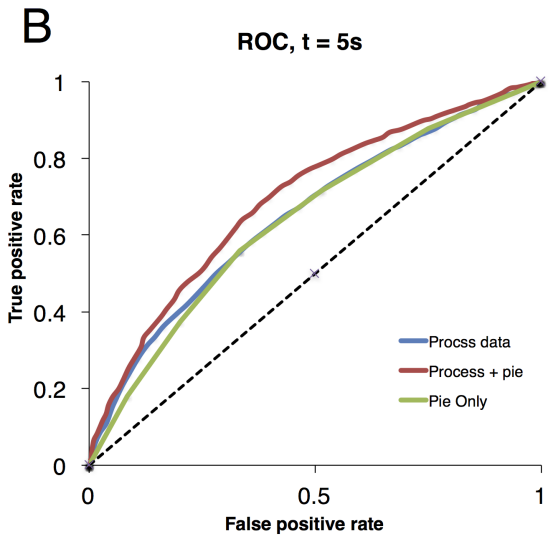

Figure 2: ROC curves showing combinations of false and true positive rates predicting bargaining disagreements. Improved forecasting is represented by curves moving to the upper left. The combination of process (cursor location features) and "pie" (amount) data are a clear improvement over either type of data alone.

Machine learning is able to find predictive value in details of how the bargaining occurs (beyond the simple, and very good, prediction from the amount being bargained over). Of course, this discovery is the beginning of the *next* step for behavioral economics. It raises questions: What variables predict? Do people understand why those variables are important? Can they constrain expression of variables that hurt their bargaining power? Can mechanisms be designed that record these variables and then create efficient mediation with no disagreement into which people will voluntarily participate?

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.

ML 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 10 useable responses. The training set consists of 70% of the observations, 30% 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 1 or 2 free parameters per person.[3]

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. 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 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.

III: Human prediction as imperfect machine learning

Some pre-history of 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 (Thaler, Misbehaving; Lewis, Undoing). It was common in those early days to distinguish anomalies about "preferences", such as mental accounting violations of

---

[3] 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}.

fungibility and reference-dependence, and anomalies about "judgment" of likelihoods and quantities.

Somewhat hidden from economists, at the time and even now, was the fact that there was active research in many areas of judgment and decision making (JDM) proceeding in parallel and conducted almost entirely in psychology departments and some business schools. JDM research was about those judgment "anomalies". 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 implicitly addressed.

An important ongoing debate in JDM 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) AI and (the judgment part of) behavioral economics.

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 50 years in the Veteran's Administration.

Meehl's mother had died when he was 16, 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 (22) which compared a set of clinical judgments with actual outcomes, and with simple linear models using observable predictors (some objective and some subjectively estimated). The domains of judgment included ____.

Meehl's idea was that these statistical models could be used as a benchmark to evaluate clinicians.[4] As Dawes and Corrigan (1974, p. 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

[4] Check if this is in Meehl "what I expected to be a floor turned out to be a ceiling"

2000) found only six in which clinicians, on average, were more accurate (and see Dawes et al 2006 Sci).

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, 1970). Kleinberg et al (2017)'s study of machine-learned and judicial detention decisions is a modern example of the same theme.

In the next couple of decades 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) gave a succinct description of this theme (p 373):

> 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.

Many of the important contributions were included in the Kahneman et al (1982) book (which in the old days was called the "blue-green bible").

In one classic study, Oskamp (1965) had eight experienced clinical psychologists, and 24 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 1.5-2 pages each about childhood, schooling, and the subject's time in the army and beyond.

The subjects had to answer 25 personality questions about the subject, each with five multiple-choice answers[5], after each of the four stages of reading. Chance guessing would be 20% accurate. Note that these questions had correct answers, based on other evidence about the case,

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

---

[5] For example, "Kidd's present attitude toward his mother is one of: (a) Love and respect for her ideals; (b) affectionate tolerance for her foibles" ; etc.

paragraph, their accuracy was 26%; after reading all five additional pages across the three stages, accuracy was 28% (an insignificant difference from 26%). However, subjective confidence in how accurate they were rose almost linearly as they read more, from 33% to 53%.[6]

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 accuracy will always increase unless fit is penalized in some way. As a result of overfitting, however, test-set accuracy can decline with more variables. The resulting gap between training- and test-set accuracy will grow, much as the overconfidence in Oskamp's subjects grew with more "variables" (more material on the single subject).

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 (labeled data, for supervised learning, in AI language). These were rarely available at the time.

Dawes (1979) proposed to give up on estimating variable weights through a criterion-optimizing "proper" procedure like OLS[7], using "improper" weights instead. An example is equal-weighting of standardized variables, which is often a very good approximation to OLS weighting.

An interesting example of improper weights is what Dawes called "bootstrapping" (a distinct usage from the concept in statistics of bootstrap resampling). The 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, 1981). Successful examples included ___.

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 which report both test-retest reliability and bootstrapping regressions indicate that only about 40% of the residual variance is unreliable noise (Camerer 1981a). 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

---

[6] 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).

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

managerial judgments improved prediction substantially over statistical models alone.

Sparsity is good for you but tastes bad

Another feature of the early statistical-vs-clinical literature was an emphasis on how small numbers of variables might fit surprisingly well.

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).[8]

For example, Dawes (1971) did a famous study about admitting students to the University of Oregon PhD program in psychology from 1964-67. He compared a three-variable model based on GRE, undergraduate GPA, and quality of the applicant's undergraduate school to an admissions committee's quantitative recommendation. The outcome variable was faculty ratings 1969. (The variables were standardized, then weighted equally.) This simple model correlated with later success in the program more highly (.48, cross-validated) than an admissions committee's quantitative recommendation (.19).[9] The bootstrapping model 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 "certain rejection" applications.

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.

---

[8] 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).

[9] 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, etc. Some of the faculty members rating those students were probably also initial raters which could generate consistency biases etc.

But people do not like to explicitly throw away information. This is particularly true if the information is already in front of us—in the form of a PhD admissions application, for example. It takes some combination of willpower, arrogance, or what have you, to simply ignore letters of recommendation, for example.

The poster child for using worthless information is personal short face-to-face interviews (___cites). There is a mountain of evidence that such interviews do not predict anything about later work performance, if interviews are untrained and do not use a structured interview format. 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, and other kinds of service and public value to get tenure several years later.

But the brain of an interviewer probably has more basic things on its mind. Is this person well-dressed? Would they protect me if there is danger? Friend or foe? Does their accent and word choice sound like mine?

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

While I do think such interviews are a waste of time[10], it is conceivable that they generate some information that is valid. The problem is that interviewers may weight the wrong information (as well as overweighting features that should be regularized to zero). Indeed, nowadays the best method to capture such information is probably to videotape the interview, combine it with other tasks resembling 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 which Kahneman and Lovallo (19_) wrote about. They called them "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" (p 25).

The outside view

"focusses on the statistics of a class of cases chosen to be similar in relevant respects to the current one" (p 25)

[10] 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.

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 as unique.

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.[11] I'll focus on two aspects: Overconfidence and how it increases; and; limited error correction.

In both cases, we are implicitly talking about a research program which collects data on human predictions and compare them with machine-learned predictions. Then *deliberately* re-do the machine learning badly (e.g., failing to correct for over-fitting) and see whether the impaired ML predictions have properties of human ones.

Overconfidence: 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.)

It could be that human overconfidence results from a failure to winnow the set of predictors (as in LASSO penalties for feature weights). Overconfidence of this type could correspond to what one expects from overfitting: High training set accuracy corresponds to confidence about predictions. A drop in accuracy from training to test means that predictions were not as accurate as hoped.

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—e.g., how children learn to recognize categories of natural objects or to learn properties of language (e.g. McClelland and Rumelhart).

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

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

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. 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 non-drums" and ignore the error. This procedure will generate too many features and will not use error-correction effectively.[12]

(Note too that a different admissions director might create two different subfeatures, "plays music in a punk band" and "plays non-punk music". In the stylized version of this description, both will become convinced that they have improved their mental model and will retain high confidence about future predictions. But their inter-rater reliability will have gone down, which puts a mathematical cap on how good average predictive accuracy can be. More on this next.)

IV: 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).

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

AI helps: A clear example is recommender systems. Good systems are a kind of behavioral prosthetic to remedy human limits on attention 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, 2015):

> 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

[12] 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.

> 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" 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, note that Netflix wants its members to "understand *why* it [a recommended video] might be of interest". This is, at bottom, a question about interpretability of AI output, how a member learns from recommender successes and errors, and whether a member "trusts" Netflix in general. All these are psychological processes which may also depend heavily on design and experience features (UD, UX).

AI 'hurts'[13]: 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 welfare implications. Behavioral economics enters with some evidence about how people make judgments about fairness of prices (e.g., Kahneman, Knetsch and Thaler, 1986[14]) and how fairness judgments influences behavior.

My intuition is that in general, people can come to accept a high degree of variation in prices for what is essentially the same product, as long as there is very minor product differentiation[15] or firms can articulate why different prices are fair (which requires insight into how consumers think, and heterogeneity). 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.

V: Conclusion

TBA

---

[13] 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.

[14] Recent?

[15] mariachi

References (partial)

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% database + 50% manager. *Management Science* 36(8). 887-899.

Camerer, C. F. (1981a). The validity and utility of expert judgment. Unpublished Ph.D. dissertation, Center for Decision Research, University of Chicago Gradu- ate Schoo] of Business.

Camerer, C. F. (1981b). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance,* 27, 411-422.

Camerer, C.F and Eric Johnson. 1991. The process-performance paradox in expert judgment: Why can experts know so much and predict so badly? " In A. Ericsson and J. Smith (Eds.), *Toward a General Theory of Expertise: Prospects and Limits*, Cambridge: Cambridge University Press, 1991.

Colin F. Camerer, Gideon Nave & Alec Smith. (2017) Dynamic unstructured bargaining with private information and deadlines: theory and experiment. Working paper

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology,* 46, 271-280.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist,* 26, 180-188.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin,* 81, 97.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science,* 243, 1668-1674.

Edwards, D. D., & Edwards, J. S. Marriage: Direct and continuous measurement. *Bulletin of the Psychonomic Society,* 1977, *10,* 187-188.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 50, 387-395.*

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemas for decision mak- ing. *Organization Behavior and Human Performance,* 13, 171-192.

Gabaix Laibson QJE

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin,* 1970, *73,* 422-432;

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt test. *Journal ofConsulting Psychol· ogy,* 23, 25-33.

Goldberg, L. R. (1968). Simple models or simple processes? *American Psychologist,* 23, 483-496.

Gomez-Uribe, Carlos and Neil 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., Zald, D. H., Lebow, B. S., Snits, B. E., & Nelson, C. E. (2000) Clinical vs. mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*:19-30.

Johnson, E. 1. (1980). Expertise in admissions judgment. Unpublished doctoral dissertation, Carnegie-Mellon University.

Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In M. T. H. Chi, R. Glaser & M. 1. Farr (Eds.), *The nature of expertise* (pp. 209-228). Hillsdale, NJ: Erlbaum.

Kahneman, Knetsch and Thaler, 1986 fairness

Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan. 2015. "The Theory is Predictive, But is It Complete? An Application to Human Perception of Randomness." Unpublished paper.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predic-tions." NBER Working Paper 23180.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review o f the evidence.* Minneapolis: University of Minnesota Press.

Mullainathan, Sendhil and Jann Spiess (2017) Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31(2): 87-106

Peysakhovich , Alexander and Naecker, Jeffrey. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. Journal of Economic Behavior & Organization Volume 133, January 2017, Pages 373-384

Sawyer, Jack (1966). Measurement and prediction, -clinical and statistical. *Psychological Bulletin,* 66, 178-200.

Thornton, B. Linear prediction of marital happiness: A replication. *Personality and Social Psychology Bulletin,* 1977, *3,* 674-676.

Daniel Kahneman and Dan Lovallo (1993)
Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking
Management Science 39:1 , 17-31

Klayman, J., & Ha, Y. (1985). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review,* 211-228.