AEA Papers and Proceedings 2025, 115: 1–23
https://doi.org/10.1257/pandp.20251118

AEA DISTINGUISHED LECTURE

Economics in the Age of Algorithms[†]

By Sendhil Mullainathan*

The rise of algorithms, particularly machine learning, does more than impact the economy. These tools offer new empirical methods. They enable new kinds of interventions. And foundational theories must be changed to account for them. Economics is, at its heart, about decisions. Because they change how we study, model, and improve decisions, algorithms should transform not just the economy but economics itself.

The title of this paper should trigger some skepticism. Understandably so. After all, if we thumb back to the very first issue of the very first economics journal (1886, *Quarterly Journal of Economics*), there was no paper titled "Economics in the Age of Electricity." Nor does that omission appear to be a missed opportunity. The logic of economics is built to weather technological innovations; we describe production at a level of abstraction that continues to apply whether factories are powered by steam or electricity. So while algorithms may affect the operation of the economy, they should not change the logic of economics—so goes the skeptical perspective.

But that skeptical perspective is wrong. Algorithms are qualitatively different from other technological innovations. They represent a change to the fundamental building block of all of economics: how decisions are made. Nearly every economic model is ultimately a model of decisions and their consequences. A radical change to how decisions are made necessitates a radical change in how economists do economics.

This argument is easiest to see in the context of a concrete example. While the example is specific, its essential features appear in many other economic decisions.

Millions of times a year in the United States, judges must make a consequential decision: Where should a just-arrested defendant await trial, at home or in jail? The stakes are high, since a long jail spell is traumatic to the defendant, while a heinous crime by a released defendant is traumatic to society. What is most relevant for our purposes is *how* the judge must decide. By law, the judge's job at this stage is not to adjudicate guilt or innocence but instead to base the release versus detain decision on an assessment of the defendant's risk of skipping court or rearrest. More precisely, judges *predict* risk.

In other words, judges are repeatedly performing a task that involves taking in a set of inputs (case and defendant facts) and using those to predict an output (risk). Behavioral economics suggests a number of reasons why that task is poorly suited for the human mind, given the need for probabilistic

* Peter de Florez Professor, Department of Economics and Department of Electrical Engineering and Computer Science, MIT (email: sendhil@mit.edu). One aspect of these lectures is antiquated: They are delivered by a single person. Economics is no longer the solo game it once was. That mismatch between how the work is done and how it is presented is especially egregious for this specific lecture. Three people in particular jointly own parts (or all) of the intellectual core here: Jens Ludwig, Ashesh Rambachan, and Jann Spiess. Still, had any one of them given the lecture, it might have been more measured and thoughtful. So I will take full credit for any excess bravado or overclaiming. I am also grateful to Janani Sekar for exceptional research assistance and to Larry Katz for selecting me to give this lecture.

† Go to https://doi.org/10.1257/pandp.20251118 to visit the article page for additional materials and author disclosure statement(s).

reasoning (e.g., Kahneman 2011; Tetlock and Gardner 2016; Thaler 2016). But notice that the structure of this problem also seems tailor-made for a supervised learning algorithm. That in turn raises a natural question: How would an algorithmic risk predictor fare relative to the judge? The answer is, it turns out to be quite a bit better (Kleinberg et al. 2018). By better prioritizing risky defendants for detention, the algorithm allows for up to a 40 percent reduction in jailing rates for the same risk level, or 25 percent reductions in risk for the same jailing rate. To be clear, these results are not an argument for automation. Instead, they illustrate how algorithms can discover information that could improve decisions. How to take advantage of that information in practice is an important question to which I return below.

These large gains naturally spur a question: What are judges doing wrong here? To answer that, in Ludwig and Mullainathan (2024), we built a new algorithm. Now, rather than predict defendant risk, we predicted judge decisions. Just as the original algorithm was a model of what the judge cared about (understanding defendant risk), the new algorithm is a model of what we as researchers care about (understanding judge decisions). And just as the original algorithm might aid the judge, this algorithm aids us as researchers by revealing some surprising facts about how judges choose. It turns out that a significant fraction of the predictable variation in judge decisions comes from a single variable: the mug shot. In fact, two features of that mug shot (learned through this process) strongly correlate with release. The magnitudes of the effect on detention are again large, larger even than that of being arrested for a violent rather than nonviolent crime. The two features this procedure identifies—well-groomed and "heavy faced"—are also novel, even to practitioners who work in the criminal justice system (e.g., a public defender's office and a legal aid society). Algorithms can thus be useful not just to the judge, but to researchers studying the judge.

This specific example illustrates several broad points.

First, while this specific example is *quite* specific indeed, other problems with the same general structure—what I call "prediction policy problems" (Kleinberg et al. 2015)—are ubiquitous. I will argue below that such problems show up over and over in every area of economics: Most of our models involve agents that (sometimes implicitly) make some prediction.

Second, algorithms expand the kinds of interventions and policy levers that economists can help design and deploy. For example, after seeing the results of our proof-of-concept algorithm, the Mayor's Office of Criminal Justice in New York City asked us to partner with some other organizations to build and implement a real-world operational version.[1] That tool is now being used every day in every pretrial courtroom in New York. Such decision aids greatly expand the scope (and potential impacts) of a category of behavioral intervention that began with nudges and choice architecture (Thaler and Sunstein 2009; Dawes, Faust and Meehl 1989).

Third, it highlights how much of the hard work that's needed going forward will be econometric, not computational. Building a supervised learner to predict risk was, relatively speaking, the easiest part of the project. The hardest part was comparing that predictor's performance to the judge's. For example, if the algorithm suggests releasing a defendant that the judge detained, we don't observe what the defendant's outcome (flight, rearrest) would have been (the "selective labels problem") (Kleinberg et al. 2018). As another example, if the judge's objective function involves components other than risk, we could mistakenly conclude that the algorithm beats the judge, but only because we're looking at too narrow a set of outcomes—so-called "omitted payoff bias." Ignoring these challenges bakes in a pro-algorithm bias. Solving these challenges requires understanding what people are optimizing, and with what information. In other words, it requires econometrics (e.g., Rambachan 2024).

Fourth, the pretrial example highlights the need for an economic framework that integrates human decision-making and algorithmic predictive capacities. We can't study how people use algorithms, nor help researchers use algorithms better, if we don't have models that formalize the comparative advantage (so to speak) of humans and algorithms.

[1] New York City Criminal Justice Agency Release Assessment (https://www.nycja.org/release-assessment)

One promising approach, in my view, emphasizes the informational advantages of each. Given a data frame and an objective function, algorithms can predict more accurately than people. On the other hand, people can see things in the world that are missing from the data frame. That suggests a useful framework for how algorithms could augment humans, not merely automate them away: human advantage in recognizing what is missing from the data, and algorithmic advantage in extracting information from the data.

Finally, the pretrial example highlights how algorithms are starting to expand the very methods we bring to science. In this case, hypothesis generation is migrating from being a "prescientific" activity to a more empirically grounded one. Since algorithms provide new tools for analyzing data and holding knowledge, they will surely lead to other substantial changes in the scientific method within economics as well, much as we have already seen in other fields (Butler et al. 2018; Jumper et al. 2021).

For economists, the rise of algorithms is currently an opportunity but will soon become a necessity. As algorithms increasingly change how *agents in the economy* make decisions, that will inevitably require changes in how we model decisions. It is, for example, no longer possible to model judge decisions in New York City without also modeling our algorithm. In this way, algorithms are following a trend. The major revolutions in economics are often revolutions in how we *understand and study decisions*. So the right analogy is not between the rise of algorithms and the rise of electricity, but instead between algorithms and game theory, information economics, behavioral economics, and the credibility revolution.

I. A Primer on Machine Learning

For economists trying to make sense of machine learning—what's new, and how it relates to the econometric tools we already know—it's easy to get lost in endless technical details. A surprisingly simple framework can help clarify what is new about these tools, where they are useful, where they are not, and as a result the kinds of economic questions to which we should apply them.

A. *Supervised Learning: Estimation versus Prediction*

Consider a simple regression:

$$Y = \beta_x X + \beta_w W + \varepsilon.$$

Sometimes we get lucky and there's enough information in the data for us to pin down the values of β_x and β_w in some economically meaningful way. But often we wind up instead with confidence intervals that include values ranging from positive to null to negative, as in Figure 1.

It is intuitively tempting to conclude that our inability to precisely estimate the parameter values β_x and β_w implies that we cannot predict Y, either. From the confidence intervals for β_x and β_w, we imagine a potential parameter space as in Figure 2, which implies an unhelpfully wide range of potential values for our predictions, $\hat{Y}$.

But as it turns out, this intuition is wrong. What this understandable intuition misses is that our explanatory variables usually *covary*. Covariance among the explanatory variables means that the confidence intervals for β_x and β_w can be consistent with a narrower set of possible (β_x, β_w) pairs. The case of positive correlation is depicted in Figure 3, which illustrates how even if we cannot confidently reject the null of zero for either, we can confidently reject the joint null. This ellipse stems from the fact that if X and W positively covary, then if the true β_x is toward the top end of its estimated confidence interval, for OLS to be minimizing a mean-squared-error loss function across $(Y, \hat{Y})$ pairs within the data frame, it cannot simultaneously be that the true β_w value is *also* toward the top end of the range of its confidence interval as well, otherwise we would wind up with a prediction that overshoots the target $(\hat{Y} > Y)$. So for a given (X, W) pair for all the coefficients in this ellipse, the predicted $\hat{Y}$ would not vary substantively.

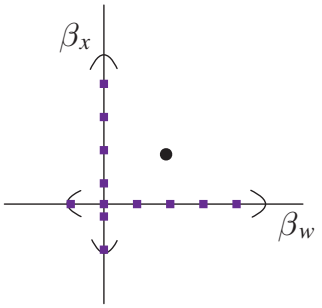

FIGURE 1.
PARAMETER ESTIMATES
WITH WIDE CONFIDENCE INTERVALS

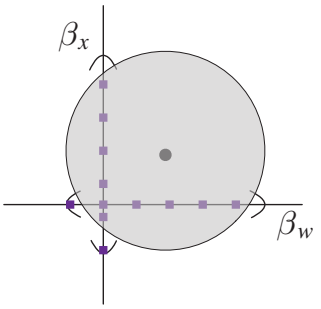

FIGURE 2.
POSSIBLE CONFIDENCE REGION

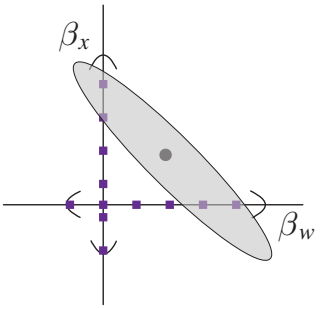

FIGURE 3.
CONFIDENCE REGION
IF X AND W COVARY

The lesson is that estimation and prediction are different; just because we can't *estimate* parameters precisely in a given application doesn't mean that we can't *predict* accurately. Estimation is about trying to apportion how the variation in Y is driven by specific explanatory variables. Prediction is about generating $\hat{Y}$ variables that tend to track closely the Y outcome variable of interest, without having to worry about which specific explanatory variable is doing most of the work. As Mullainathan and Spiess (2017) put it, estimation is about solving $\hat{\beta}$ problems, while prediction is about solving $\hat{Y}$ problems.

I see three particularly useful insights that flow from this simple framework.

First, it helps economists understand at what level of abstraction they need to engage with the computer science research literature on machine learning. It's easy to get overwhelmed by the firehose of specific details about each potential algorithmic function class: LASSO, ridge regression, support vector machines, neural networks, decision trees, gradient-boosted trees, random forest, etc. From the economist's perspective, one could abstract from those implementations and instead view them all as just tools for generating $\hat{Y}$ predictions.

For most economic applications that's enough. Imagine we try to see if judges mispredict risk by comparing the predictions implicit in their release versus detain decisions with those of an algorithm. The specific algorithmic function class used for that prediction is typically a second-order issue for both a practical reason and a conceptual reason. Practically, it turns out that the amount of data brought to bear (both observations and variables) is typically more important than the specific function class used. And conceptually, if even just a subset of function classes outperformed the judge, we have still learned something important about judge decision-making.

Second, the simple regression example illustrates why machine learning can handle high-dimensional datasets (including more variables than observations) and highly flexible functional forms. Imagine, for example, a dataset of news stories about different companies connected to stock price information. There would be far more total words (variables) than total news stories (observations). That means that multiple models could be built with the dataset that would fit the stock price outcome equally well. How do we tell which specific model is right? Answering that difficult question is the key goal of estimation problems. For prediction problems, machine learning solves that problem by simply not caring about it. Machine learning asks instead: Which of these candidate models predicts most accurately?

Third, and relatedly, this simple framework highlights how machine learning and standard econometrics relate to one another. When giving talks about machine learning I am often asked, "So is machine learning simply *better* than traditional econometric tools?" But that's the wrong question to ask. The right way to think about machine learning relative to econometrics is that they're designed to do different things. If we care about inferences on parameters ($\hat{\beta}$ problems), we need traditional econometric estimators. If we care about inferences about prediction quality ($\hat{Y}$ problems), machine learning tools are far more suitable. They're different tools for different tasks.

B. *New Tools, Old Problems versus New Tools, New Problems*

Once we realize that machine learning is well-suited for solving $\hat{Y}$ problems, the natural next question to ask is: So, where are the $\hat{Y}$ problems in economics?

New innovations often get applied first to old problems. In economics, that's taken the form of repurposing machine learning $\hat{Y}$ tools to solve traditional $\hat{\beta}$ problems. For example, can machine learning help choose control variables in a regression to minimize residual variation in the outcome (Chernozhukov et al. 2018)? Can we use machine learning to estimate conditional average treatment effects (Athey and Imbens 2016; Wager and Athey 2018; Künzel et al. 2019)? Can we use machine learning to incorporate new high-dimensional data sources like text, images, cell phone GPS coordinates, etc. into more traditional estimators like regression (see, e.g., Blumenstock, Cadamuro, and On 2015; Jean et al. 2016; Gentzkow, Kelly, and Taddy 2019; Dell et al. 2023)?

With time, though, new innovations often reveal new problems. I will focus here on two broad categories of important $\hat{Y}$ problems that show up broadly across economics.

The first involves decisions made by economic agents out in the world that hinge not on a causal inference but on a predictive inference instead. Judges deciding which defendants to release is a canonical example but far from the only example. Machine learning enables the positive research activity of studying how those decisions are made and the normative research activity of trying to improve them. We call these "prediction policy problems" (Kleinberg et al. 2015).

The second type of decision seems very different on its face but winds up having exactly the same essential structure: not decisions made by economic agents in the world whom economists wish to study but rather decisions by economists themselves in the act of economics research. Examples include coming up with anomalies that tell us that existing theories are inadequate, and generating entirely new hypotheses to test.

Both types of problems look quite different from those that economists normally work on—decisions that hinge on a causal inference rather than a predictive one. But, as I will argue next, I think these prediction decisions, both out in the economy and by economists trying to understand that economy, are at least as consequential.

II. Prediction Policy Problems[2]

Countless times every day all around the world, people are making decisions that hinge on a causal inference: policymakers choosing what sort of job training to offer unemployed workers, doctors deciding what sort of chemotherapy to give to a patient, judges deciding whether electronic monitoring is capable of reducing pretrial recidivism without having to resort to jailing people, teachers trying to figure out how best to remediate a student's academic deficiencies, etc. No wonder economists spend so much of their time studying these decisions.

But there are many, many other decisions made countless times all over the world that have a *different* structure; they hinge not on a causal inference, but rather on a predictive one. In what follows, I will talk about what distinguishes these two types of decisions, and why I think solving what I call "prediction policy problems" is at least as scientifically interesting and socially important as solving causal inference problems.

A. *Causation versus Prediction Problems*

Let me start by returning to the pretrial example introduced above. The judge has to decide whether to detain the defendant or not (D). The law says the judge is supposed to focus on the risk the defendant skips court or reoffends (Y). So the judge makes that decision by forming a prediction of the defendant's risk, using some defendant characteristics that are captured in data (like current charge

[2] This discussion draws from Kleinberg et al. (2015).

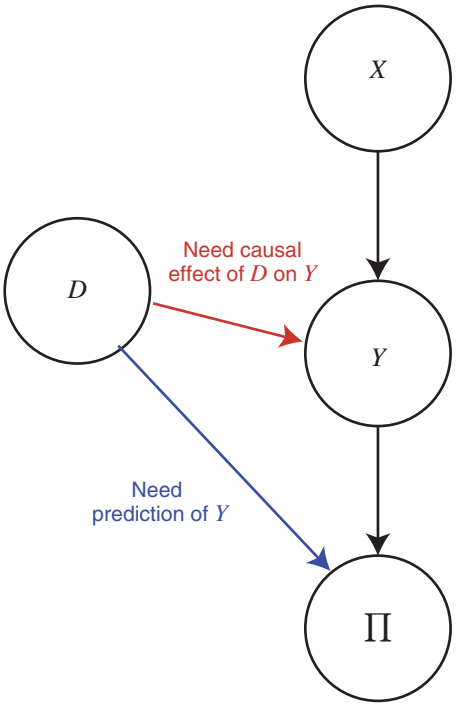

FIGURE 4.
PREDICTION AND CAUSATION PROBLEMS

and prior record captured by the rap sheet), X, as well as some that are not, Z (such as what is said in court), or $h(X,Z)$. Let the true underlying relationship given by nature be $f(X,Z)$.

One key feature of this decision is that the defendant's risk is the defendant's risk. That is, the judge's decision about whether to detain the defendant or not doesn't change that underlying risk. Instead, the judge's decision simply determines whether society releases the defendant and takes a draw from that risk distribution. That is, the decision determines how the defendant's underlying risk translates into social welfare, or

$$\Pi\big(f(X,Z),X,Z,D\big).$$

Contrast this with a decision that hinges on a causal inference instead, like whether the judge should allocate some intervention to a defendant like electronic monitoring (or drug treatment or mentoring or GED classes or a job, etc.). The key criterion on which the decision hinges is the answer to the question "What would this do to someone's recidivism risk?" That is, the decision aspires to directly causally influence the realized value of the outcome Y.

There are two ways to see the distinction between causation and prediction. First, Figure 4 shows a causal graph. When the decision has a (potential) effect on Y, then we need a causal estimate. When the decision instead only affects payoffs Π, then we need to predict Y.

One can also see this through an equation by forming a total derivative of payoffs:

$$\frac{d\Pi}{dD} = \left[\frac{\partial \Pi}{\partial D}\right](Y) + \left[\frac{\partial \Pi}{\partial Y}\right]\left[\frac{\partial Y}{\partial D}\right].$$

The first term on the right-hand side says that the decision's impact on utility depends on the value of Y. That's prediction, where the key object of empirical interest is $E\left[Y|X,Z\right]$. The second term on the right-hand side reveals that a second way in which a decision could affect utility is if it changes the value of Y itself, which then changes utility or welfare. That's causal inference, where the statistical object that must be estimated is equal to $E\left[Y|D\ =\ 1\right] - E\left[Y|D\ =\ 0\right]$.

Economists have not paid much attention to prediction problems historically in part because our empirical tools are geared more toward causal inference than prediction. But just because some of the relevant empirical tools for prediction problems come out of computer science doesn't mean that economists don't have an important role to play in solving such problems. One reason is that while *building* the algorithm can in some cases be a mechanical application of known computer science techniques, *evaluating* the algorithm's performance is far from straightforward.

How do we know whether the algorithm's predictions are better than those implicit in the judge's release decisions? When the two disagree about a case the judge releases—the algorithm claims they are actually high risk, not low risk—we can figure out who's right by looking at whether the defendant reoffends or skips court. But what if the algorithm's prediction disagrees for a case the judge detained—the algorithm claims they were actually low risk? We can't directly observe the counterfactual of what the defendant would have done had they been released, the "selective labels" problem (Kleinberg et al. 2018).

A common solution is to assume that judge decisions are random conditional on observable defendant characteristics. The researcher in that case simply imputes the missing flight or reoffending outcome for jailed defendants by using data on the defendants whom the judge released. But that approach basically assumes the answer to the question of whether the algorithm is better than the judge, since it assumes away the one potential advantage the judge has over the algorithm—the judge sees things about each defendant that the algorithm does not (Z).

Plus, there's another challenge sitting in the background: How do we know that the outcome the algorithm is predicting is really what the judge is trying to optimize? Do judges prioritize the prediction of someone's risk for serious violent crimes, like murder or robbery, or is the judge really worried about reoffending for any sort of crime at all? (The overwhelming majority of crimes are low-level offenses.) Or do judges pay attention to mitigating racial or other disparities in whom they detain? Or do they put extra weight on avoiding detention for someone who has a job or is a caretaker to one or more children? Mis-specifying the judge's objective function creates what Kleinberg et al. (2018) call "omitted payoff bias."

Notice that ignoring either or both problems—selective labels and omitted payoff bias—winds up baking in a pro-algorithm bias to any evaluation. The only solution is to carefully specify the decision-maker's objective function and come up with some way of credibly identifying the counterfactual outcomes of defendants whom the judges detained. That's not advanced computer science. But that *is* advanced economics (Kleinberg et al. 2018; Rambachan 2024).

B. *Prediction Problems Are Ubiquitous*

While I have used a concrete example of judge pretrial decisions to illustrate key features of prediction policy problems, such problems show up in almost every domain economists care about. Prediction is often the hidden flip side of specific decisions economists have been studying for decades, or lurking around within domains we've been working on for years but have missed because of our focus on causal inference problems, or even sitting there within entire domains we've largely ignored because it wasn't even obvious what work could be done within them.

While Table 1 provides a long list of examples, let me briefly discuss a few illustrative ones.

For decades, economists and other social scientists have studied and argued about the decision of whether to hire more teachers to reduce class sizes to potentially raise student learning.[3] We have

[3] See, for example, Coleman (1968); Jencks et al. (1972); Krueger (1999, 2003); Hanushek (1999); and Jackson, Johnson, and Persico (2015).

Table 1—Examples Illustrating the Breadth of Prediction Policy Problems in Economics

Economic domain	Concrete example
Decisions we already study but ignore prediction component	
Education	We study the causal effects of hiring more teachers to reduce class size (Krueger 1999, 2003; Hanushek 1999) but ignore the selection of which teachers to hire (Chalfin et al. 2016).
	We study the determinants of how students succeed in their classes in school but ignore student selection of those classes (Bergman, Kopko, and Rodriguez 2021).
Health	We study doctor treatment decisions for moral hazard but ignore potential doctor misprediction of patient health (Abaluck et al. 2016; Currie and MacLeod 2017; Mullainathan and Obermeyer 2022; Daysal et al. 2022).
Domains we study but ignore prediction-dependent decisions	
Finance	Which borrowers will repay their loans? (Rambachan, Coston, and Kennedy 2024; Blattner and Nelson 2021)
	Which people will perform well serving on a corporate board? (Erel et al. 2021)
Public finance	Which income tax returns should be audited? (Battaglini et al. 2024; Elzayn et al. 2025)
Labor	How do firms hire workers, and how could they do better? (Li, Raymond, and Bergman 2024)
	How do job seekers pick where to apply, and how do they predict offer likelihood? (Behaghel et al. 2024; Le Barbanchon, Ubfal, and Araya 2023)
Crime/criminal justice	How do police allocate resources based on predictions about where and when crime occurs? (Mohler et al. 2015; Jabri 2021)
Domains we largely ignore because it's not obvious what we could even study	
Child welfare	Which child abuse reports should be investigated? (Chouldechova et al. 2018; Grimon and Mills 2025)
Humanitarian aid	How can we target aid to where and when it will be most needed? (Aiken et al. 2022; Callen et al. 2024)

ignored the prediction-related flip side of this question: *Which* teachers should we hire? That depends on the predicted productivity of each applicant (Chalfin et al. 2016). Note the causal effect of more teachers *cannot* be separated from how we pick teachers, since the effect of those marginal teachers depends critically on their average quality.

Or consider the case of doctor decisions about which patients to refer for additional testing. Economists have focused on the causal effects of the incentives facing doctors. Evidence that many tests yield a low prevalence of discovered health problems is interpreted as a sign of moral hazard. That interpretation ignores an upstream decision by doctors about how they rank order patients by predicted risk. The implicit assumption of the moral hazard interpretation is that doctors are, for personal financial reasons, simply setting the testing threshold too low in the predicted risk distribution. But machine learning helps us see that assumption winds up being incorrect; low testing yield occurs in large part because doctors are mispredicting patient risk and testing too many low-risk patients while not testing many high-risk ones (e.g., Mullainathan and Obermeyer 2022).

These are examples where economists were already focused on a decision; they just missed the prediction component. There are also countless other examples of economists ignoring a prediction-related decision altogether, even though we were already working in that broad domain. For example, many public finance papers estimate how people respond to changes in marginal tax rates. Yet an adjacent decision is: How should we enforce those tax rules? Which returns should we audit? That's a prediction policy problem: Which audit is most likely to be fraudulent (Battaglini et al. 2024; Elzayn et al. 2025)? This enforcement decision is not only scientifically important, since it must surely affect people's behavioral responses to changes in tax rules, but also substantively important: The Internal Revenue Service estimates that something like $600 billion in owed taxes goes unpaid each year in the United States.[4]

[4] https://www.pgpf.org/article/the-united-states-forgoes-hundreds-of-billions-of-dollars-each-year-due-to-unpaid-taxes/

A third category of prediction policy problem shows up in domains that economists have largely ignored historically, partly because it has simply not been obvious what there is for us to do. Consider, for example, child welfare. This is an important domain where, it seems fair to say, there is not an excessive amount of work by economists.[5] There may or may not be a lot of economics questions here related to incentives, but this is a domain where there are important prediction policy problems like: Which child abuse reports should be investigated? Recognizing that we can study (and potentially improve) those decisions with algorithms opens the door to a whole new area of research for economists (see, e.g., Grimon and Mills 2025).

Prediction problems are not just ubiquitous in the sense of "present in every domain economists care about," but also in terms of their sheer volume. Something like ten million people are arrested every year in the United States. About four million people change jobs every *month*.[6] The United States has 54 million children enrolled in K–12 at any point in time and another 18 million college students. CDC estimates that there are something like one billion (with a "b") doctor-patient interactions every year[7] in America's $4.9 trillion (with a "t") health care sector (equal to nearly 18 percent of GDP[8]). The frequency—and often high stakes—of decisions hinging on a prediction suggests that the potential social welfare gains from understanding and improving them could be enormous. That is the point to which I turn next.

C. *Unreasonable Cost Effectiveness*

Hendren and Sprung-Keyser (2020) introduce the idea of the "marginal value of public funds" (MVPF) to provide a unified social welfare analysis of 133 historical policy challenges. The MVPF essentially divides the benefit to society by a policy's net cost to the government. Their review of past policies finds that 19 out of 133 yield an MVPF of infinity (the policy's net cost to government is zero, so social benefits are possible with no additional public spending).

By comparison, we looked at four different algorithmic policy interventions—for pretrial release, selecting college courses, heart attack screening, and reducing worksite injuries—and each of the four has MVPF of infinity, as shown in Figure 5 (Ludwig, Mullainathan, and Rambachan 2024). For example, in the case of pretrial judge decisions, our policy simulation suggested that it might be possible to reduce crime by 25 percent without increasing jail populations or, alternatively, reduce the jail population by 40 percent with no increase in crime (Kleinberg et al. 2018). The size of these changes (and the large reductions in various government costs that would result) in relation to the relatively modest cost of building and deploying an algorithmic decision aid for judges is partly what motivated the Mayor's Office of Criminal Justice in New York City to ask our team to help build and implement the real-world version.

These large social welfare gains are possible with algorithmic interventions for three reasons, two of which are obvious and one of which is less so.

The first obvious explanation is the sheer scale of impact that algorithms can have. "Normal" policies typically show diminishing marginal returns, both because implementation fidelity can be hard to maintain as scale increases and because of (if a program relies on inelastically supplied inputs) rising marginal costs (Davis et al. 2017; List 2022). As President Bill Clinton put it, "Nearly every problem has been solved by someone, somewhere … we can't seem to replicate [those solutions] anywhere else." [9] But software, in contrast, runs perfectly over and over again at almost zero marginal cost. That, combined with the huge volume of prediction-related decisions made every year, allows algorithms to have impact at such a large scale.

[5] There are of course some important exceptions; see, for example, the excellent work in Doyle (2007, 2008) and Bald et al. (2019).

[6] https://www.pewresearch.org/social-trends/2022/07/28/majority-of-u-s-workers-changing-jobs-are-seeing-real-wage-gains/

[7] https://www.cdc.gov/nchs/fastats/physician-visits.htm

[8] https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet

[9] https://ssir.org/articles/entry/going_to_scale

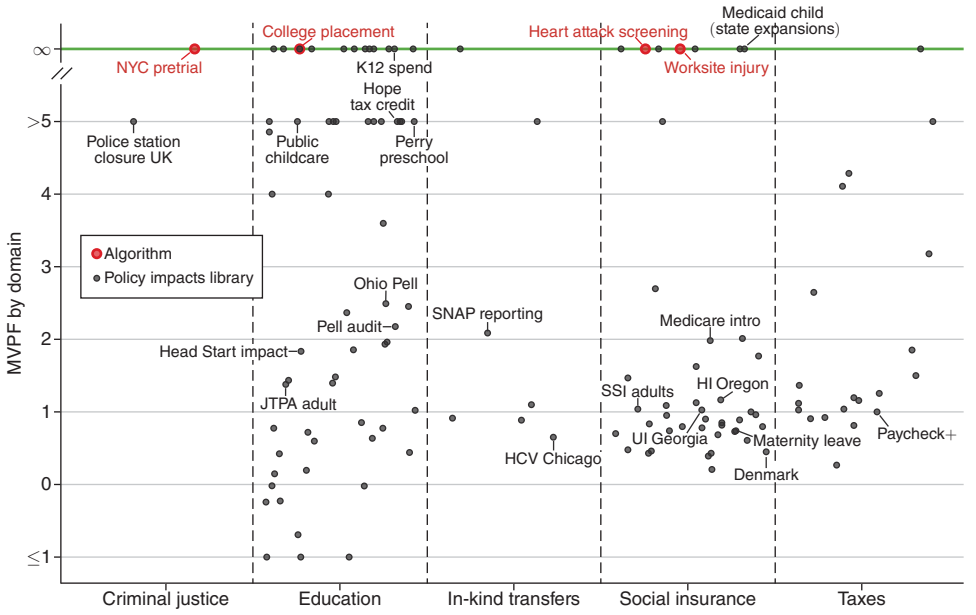

FIGURE 5.
COMPARING MVPF VALUES FOR TRADITIONAL POLICIES TO THOSE OF ALGORITHMS

Source: Ludwig, Mullainathan, and Rambachan (2024).

The second obvious explanation is that the alternative to the algorithm—a person—leaves so much room for improvement. An enormous body of research in psychology and behavioral economics documents all the ways in which people make mistakes, even with very high-stakes decisions. For example: We rely on biases and heuristics (e.g., Kahneman 2011; Thaler 2016), such as relying too much on examples that come readily to mind (Tversky and Kahneman 1973), ignoring base rates (Nisbett and Borgida 1975), or overgeneralizing from tiny snippets of information (Tversky and Kahneman 1971); we cling to our priors too tightly (Tetlock and Gardner 2016); or our cognition suffers from too much inconsistency or "noise" (Kahneman, Sibony, and Sunstein 2021). Algorithms, in contrast, can help avoid many of these errors.

The third, less obvious reason that algorithms can have so much impact stems from the logic of ranking problems at the heart of so many economically important decisions. The usual logic of policy interventions is that the government has properly rank ordered cases by marginal benefits, and so expansions of the policy serve additional cases with relatively lower benefits (Figure 6). But government agencies and decision-makers regularly misrank. The algorithm can improve the rank ordering of cases, yielding a steeper social returns schedule and a sizable reduction in deadweight loss between the flatter and steeper schedules (Figure 7).

D. *Open Challenges*

So what's stopping us from realizing such large gains?

For most of the economically important prediction problems out in the world, society has expressed a strong preference for algorithms to serve as decision aids, not decision-makers. People are comfortable with a judge looking at an algorithmic decision aid, but not with an algorithm that replaces a judge. The large body of evidence showing that *on average* statistical models predict more accurately than humans (see, e.g., Ohlin and Duncan 1949; Meehl 1954; Dawes, Faust, and Meehl 1989;

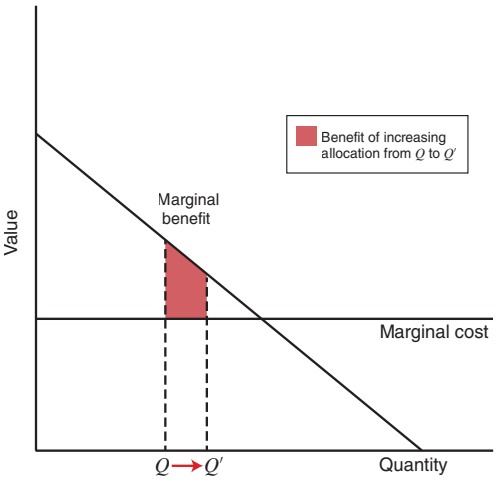

FIGURE 6.
STYLIZED ILLUSTRATION OF THE SOCIAL WELFARE GAINS IN INCREASING ALLOCATION AT CURRENT RANKING

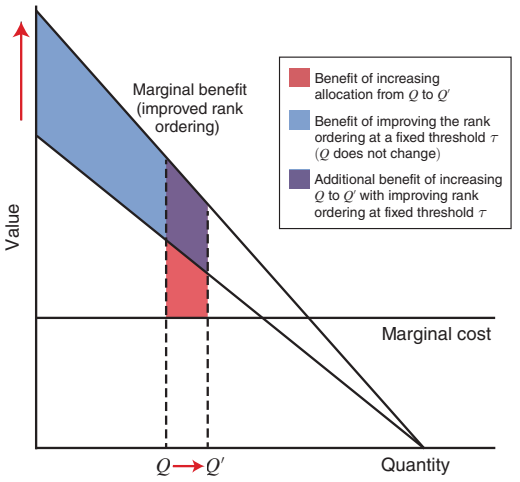

FIGURE 7.
STYLIZED ILLUSTRATION OF THE SOCIAL WELFARE GAINS FROM ALGORITHMIC RERANKING OF WHO IS PRIORITIZED FOR SERVICES

Kleinberg et al. 2018; Rambachan 2024) leads to the understandable intuition that we should simply try to get humans to follow the statistical model's recommendations as often as possible. But it turns out that people don't seem to love mindlessly doing everything a decision aid says. That's led to a growing body of research on understanding the reasons for this "algorithm aversion" (Dietvorst, Simmons, and Massey 2014, 2015, 2018; Dietvorst and Bharti 2020; Burton, Stein, and Jensen 2020; Jussupow, Benbasat, and Heinzl 2020).

The barrier to progress here is that we're missing a conceptual framework that combines what the human decision-maker sees and what the algorithm sees. To see what I mean, in what follows I sketch out an initial example of what such a framework might look like using our running example of judge pretrial decisions (see also Ludwig and Mullainathan 2021).

As noted above, the judge's decision about whether to release versus detain a defendant pretrial is by law supposed to hinge on a prediction of the defendant's flight or public safety risk. One way we could learn the relationship between the defendant's outcomes if released and their observed and unobserved characteristics, $Y = f(X,Z)$, is with an algorithm, which by construction can only make use of whatever variables are actually captured in data, $E\left[f(X,Z) \mid X\right]$. The algorithm can predict accurately for a given data frame, but the algorithm has the drawback of not having access to all the potentially relevant information. The algorithm can only predict using the Xs.

Alternatively, a human could form that prediction. A large body of behavioral science research shows that most people have enormous difficulty predicting accurately. Let the human predictions be given by $h(X,Z) = f(X,Z) + \Delta$. This pattern of misprediction is what naturally leads to the intuition that the algorithm is surely better. But notice that the human also has two potential advantages relative to the algorithm. First, the human has access to more information than the algorithm does. The judge can predict using things not captured in the rap sheet data—that is, the "unobservables," Z, like what's said in the courtroom, etc. Second, unlike the algorithm, people know what they (the human decision-makers) are trying to optimize, Π.

Notice some implications that flow from even such a simple placeholder framework as this.

First, this framework helps us see why people *shouldn't* just mindlessly follow an algorithm's predictions and recommendations: because they (the humans) also have their own sources of comparative advantage relative to the algorithm. Conceptually, it is possible that even though the algorithm does better than the human *on average*, in principle there may be at least a *subset* of cases where the

human could be doing better. Put differently, the framework makes clear why in principle the human and machine together could potentially do better than either alone.

Second, this framework captures the evaluation challenges described above, selective labels and omitted payoff bias. The judge makes release decisions based on $h(X,Z)$, so we can't simply impute the missing outcomes using the algorithm's predictions $m(X)$. And to determine whether the algorithm outperforms the judge, we don't really want to compare the predictions $m(X)$ and $h(X,Z)$, but instead we want to compare $\Pi(m(X),X,Z,D)$ with $\Pi(h(X,Z),X,Z,D)$. How do we do that in a world in which we don't directly observe social welfare or utility? That's a classic economics problem.

Finally, this framework clarifies an open scientific problem: understanding the design space of possible decision aids. It is tempting to imagine that the only option is to show the decision-maker the algorithm's predictions. But one could do so much more. The algorithm could simply serve as a mirror on the judge's past choices. Rather than giving suggestions on particular choices, it could allow decision-makers to explore what has worked and what has not in the past. That could help them learn strengths and weaknesses of their own decision-making. Alternatively, it could simply highlight in any particular case aspects that might lead the decision-maker astray. For example, it could have pointed out, "This defendant is not well-groomed, and in those cases …" Or the algorithm could simply sort what order decisions are seen in, so that the toughest cases are prioritized for more time or mental bandwidth. Or is the right solution something else entirely? This isn't just a hard problem; it's a problem that involves both humans and machines together. So by its nature it cannot be solved by computer scientists alone; it must inevitably involve economists as well.

III. Machine Learning for Science

So far I have discussed prediction policy problems like judges looking at data on defendants to decide whom to release or detain, doctors looking at data to decide which patients to refer for more testing, HR managers looking at data on past hires to decide which applicants would be most productive if hired, etc. Economists can then use data to study and potentially even improve those decisions. I turn next to a different but conceptually related problem, involving an economic agent looking at data to make a decision that hinges on a prediction: economists carrying out the scientific enterprise of economics.

A. *Machine Learning for the Natural Sciences*

To see how algorithms might help with economic science, let's start with how and why they've been so useful for the natural sciences.

Consider the problem of protein folding. In that scientific application, we know we have the correct theory of how proteins fold. The problem is not one of inadequate theory but rather one of computing; empirically understanding the implications of the theory for different specific compounds is simply too complex. Lab experiments are possible but costly. Machine learning in this case has been enormously valuable in turning theory into something tractable. The scientific gains here are not small. The Nobel Prize in chemistry was recently awarded for the protein folding application (e.g., Jumper et al. 2021), with similarly impressive gains for the discovery of inorganic crystals (Merchant et al. 2023) and solving challenging problems in mathematics (e.g., Romera-Paredes et al. 2024).

But for the economic sciences, because our theories are in a different place, the role of algorithms in the scientific enterprise must be different as well.

B. *Algorithms for "Incomplete" Sciences Like Economics*

Unlike with the natural sciences, our theories in economics are not correct. They might have some predictive power, but there's a *lot* that they can't explain—they're "incomplete" (Fudenberg et al. 2022). So the role of the algorithm isn't about making a correct theory tractable; it's instead about how to make incomplete theories better.

Thinking about theory completeness made clear a curious asymmetry in how economics is done. For hypothesis testing, we have elaborate methods for everything we do ranging from causal identification (Angrist and Pischke 2010) to statistical inference (e.g., Bertrand, Duflo, and Mullainathan 2004; Kling, Liebman, and Katz 2007) to generalizability (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996). But for hypothesis generation, for the process of coming up with the ideas we test in the first place, we have . . . what exactly? That process is remarkably informal, idiosyncratic, and ad hoc.

Algorithms have the potential to improve and accelerate the process of theory generation to complement what humans are already capable of doing. Consider, for example, the problem of choice under uncertainty. We first had Von Neumann and Morgenstern's expected utility theory. To push the science forward, researchers worked hard to identify pairs of examples that are both consistent with the current theory but inconsistent with one another—so-called "anomalies." This type of activity led to findings like the Allais paradox, the common ratio effect, and various Kahneman and Tversky anomalies. Those led to a new theory, prospect theory, followed by new anomalies, leading to yet new theories like salience theory (Bordalo, Gennaioli, and Shleifer 2012), simplicity preferences (Oprea 2022; Puri 2025), cognitive uncertainty (Enke and Graeber 2023), etc.[10]

Anomaly generation is a powerful way to improve theories, but slow. How might algorithms help scientists accelerate that process? Mullainathan and Rambachan (2024) illustrate one way to do this. Given a dataset and given a theory, the algorithmic anomaly-generation procedure works by converting the problem into max-min optimization, where a theory seeks to maximize its explanatory power over a dataset by minimizing the loss between the set of allowable functions under the theory and the observed data while a "falsifier" algorithm generates candidate datasets to find data that maximize the loss of the theory—that is, a dataset that misfits the theory. The procedure, given expected utility theory and data on lottery choices, winds up on its own rediscovering the known anomalies for expected utility theory; moreover, it suggests new anomalies not in the existing economics literature. New data collection on these lottery menus confirms that they are indeed expected utility violations.

Anomaly generation works well for building on explicit theories we already have, but what do we do in cases where there's a lot going on out in the world that we simply have no way to currently explain? Consider, for example, judge pretrial decisions. Kleinberg et al. (2018) show that judges seem to substantially mispredict defendant risk. But what exactly are judges getting wrong?

To see how algorithms can help aid scientists in answering this question, we need a conceptual framework that helps us integrate both what the algorithm sees and what the scientist sees. It turns out that that framework looks a lot like the one for prediction policy problems.

For starters, let's consider what human scientists currently do. In the case of pretrial hearings, the scientist comes up with a hypothesis for what the judge might be getting wrong by forming a candidate model of how the judge might be making their decisions, $h(X,Z)$. The scientist does that on the basis of what's in the available data (the Xs), which in this case would include, for instance, the rap sheet and other information about each defendant and, ideally, also information about each judge who is hearing pretrial cases. The scientist also draws on information that is not captured by any dataset, which could include personal observations about the world as well as theories, hunches, etc., some of which may be tacit—that is, the scientist themselves may "know" it even if they are not necessarily able to consciously articulate it, a version of what Autor (2014) calls "Polanyi's paradox." For example, scientists know that racial bias shows up in lots of decisions out in the world and, on that basis, hypothesize that there could be bias in judges' pretrial decisions as well (see, e.g., Arnold, Dobbie, and Yang 2018).

Alternatively, we could use an algorithm to model judicial decisions instead, $m(X)$, and then form hypotheses about what the judge might be doing from looking at that prediction function. For example, comparing an algorithm built to predict the judge's decisions with an algorithm built to predict

[10] Relatedly, there is also the wonderful paper by Harless and Camerer (1994) that sits above this anomaly generation and theory development process and compares how well different theories can explain these anomalies.

FIGURE 8.
EXAMPLE OF MORPHED SYNTHETIC FACE

the defendant's risk directly suggests that judges overweight the current charge for which someone was arrested (Sunstein 2021).

One immediately obvious implication of even this simple framework is that a new skill set will be required by economists interested in using algorithms for hypothesis generation. Our usual research activity of causal testing of hypotheses is essentially all about narrowing down the data—finding plausible sources of identifying variation that helps the economist find two comparable groups to compare. But *generating* hypotheses is all about *expanding* the data—what additional data sources could be brought to bear to shed new light on the judge's behavior? That is, the algorithm's model of the judge—and hence the set of new hypotheses the algorithm can identify—is constrained by those data features that the scientist thinks to give the algorithm access to, X. For example, Eren and Mocan (2018) came up with the clever (human-generated) hypothesis that judges are affected by their mood, so that even trivial events like whether the big state university football team won or lost last weekend could affect high-stakes pretrial decisions. An algorithm could only have come up with that hypothesis on its own only if some clever scientist thought to merge court records with massive archives of general news stories.

As with prediction policy problems, this simple framework helps us see why for hypothesis generation there is the potential for scientists working with algorithms to do better than either on their own. The human has access to things the algorithm does not, but can mispredict, so that if the "true" model of the judge's behavior is $f(X,Z)$, the scientist's mental model of the judge is $h(X,Z) = f(X,Z) + \Delta$. The algorithm predicts accurately but can only model the judge's behavior using things that are in data, so $m(X) = E\left[f(Y,Z) \mid X\right]$.

One example of how to solve this problem is in Ludwig and Mullainathan (2024), where we build an algorithmic model of the judge bringing as much data to bear as possible, including high-dimensional data sources like images. That statistical model capitalizes on the great strength of the algorithm relative to the human: to predict the judge's behavior as accurately as possible or, said differently, to notice patterns in the judge's decisions that the scientist does not. As noted in the introduction, we find that one of the strongest predictors of judge decisions is something that is surely irrelevant to what the law asks the judge to consider (defendant risk): the defendant's face.

The reason that the algorithm's prediction $m(X)$ is not itself a useful hypothesis is because it is not interpretable. The neural network we built using the defendant's rap sheet, mug shot, and other information has billions of parameters and lots of complicated interactivity among data features. Knowing *something* about the defendant's face doesn't give scientists the guidance they need to carry out what is perhaps the most important thing that pushes science forward: to figure out what new things to measure. That is, historically one of the key scientific activities has been to decide which currently unobserved things, Z, to observe and turn into data, X.

Our solution was to build a hypothesis-generation pipeline that enables the algorithm to communicate the signal it has discovered to humans by morphing pictures of defendants along the predicted-risk gradient in the direction of higher detention odds. Human study subjects then look at the morphed image pairs and articulate what defendant features they see as changing, since one of the key things that humans know and algorithms do not is what concepts human beings can understand. See Figure 8. That has led to several new hypotheses about judge errors that open the door to new scientific work in this area and, potentially, new policy interventions as well.

These papers are merely the first word on these issues. Capitalizing on the power of algorithms to help economists do science will require solving a variety of new methodological questions. These

questions aren't just missing answers; many of the key questions haven't even been asked yet. The first step is of course tool building; the anomaly-generation and hypothesis-generation tools described here are examples, as are measures of completeness. There are surely many others. The second step is frameworks to help analyze properties of these tools. Compared to hypothesis testing, we have very little if any formal frameworks within which to analyze these tools. For example, with hypothesis testing, we have excellent frameworks to think about things like power and size; what are the equivalents for hypothesis generation? What's already been done barely scratches the surface of what's needed.

IV. From Opportunity to Necessity

If we look backward, we can see that algorithms have created important new research opportunities for economists. Looking forward, it seems clear that algorithms in economics will go from opportunity to necessity. The reason is that economic agents themselves will increasingly be using algorithms for countless decisions made in the real-world economy. As the agents in our models change the way they make decisions, the existing models will become outmoded and we will need new ones.

Consider the example of an unemployed worker searching for a new job. To which job openings should they apply?

The positive study of this question involves understanding how the worker optimizes given utility and information. The worker makes choices based on their preferences Π and a prediction of whether they will get any given job if they apply for it, $\hat{Y} = h(X,Z)$. The prediction is made by the worker reflecting on their own "quality," which is related partly to things captured in data that firms can see, like schooling attainment $(X\text{s})$, and some of which is not, such as reliability, motivation, etc. $(Z\text{s})$. The key preference parameter that the worker knows about themselves is their reservation wage. Together these factors shape the decision about which jobs to apply to and how many—that is, how much effort to expend on job search.

The normative study of this question includes the design of unemployment insurance (UI) benefits. In designing that benefit regime, the goal is to balance the desire to provide people with social insurance against the potential for moral hazard. The key parameter of interest for how to balance these opposing forces is the elasticity of search behavior with respect to program generosity, η, which is usually treated as a sufficient statistic.

How can an algorithm help here?

Since job search is a prediction problem and humans make those predictions imperfectly on their own, $h(X,Z) = f(X,Z) + \Delta$, an algorithmic decision aid has the potential to improve job search decisions and outcomes. Recent work suggests that there could be large gains (Behaghel et al. 2024).

But notice what happens once an algorithm is introduced into the job search process. *It is no longer possible for economists to model job search by simply modeling the worker*. Job search is now the joint outcome of the worker plus the algorithm.

Notice also a profound implication of evidence that the introduction of such an algorithm changed people's search behavior: It would mean that our previous models of this behavior had been wrong. That is, if access to an algorithm changed job search so that the observed elasticity η changed, it could not have been the case that the initial value of η that economists had estimated from the pre-algorithm data was reflective of the worker's true underlying preferences. That previous estimate was in part also reflecting worker error. Our existing model of the worker's job search decision *must* have been wrong.

Not only does our scientific understanding of positive economic questions change, but our efforts on normative questions like "How can we improve the unemployment insurance system?" must also change. The reason is because it is no longer possible to think only about what the *worker* will do. If the value of η depends on the algorithm, then this key behavioral elasticity is not an exogenous parameter to estimate but rather something endogenous to the policymaking process. The implication is that policy design has no choice but to jointly optimize not just the UI benefit schedule but the algorithm as well. There is simply no way to abstract away from the job search recommender system.

Does this mean that economists must be in the algorithm design business? Yes. But isn't that really the role of computer science? Absolutely not. Notice what the recommender algorithm is doing for

workers: making recommendations from what the data say about the experiences of similar workers in the past. Job openings are then rank ordered based on the probability that a similar worker in the past applied to such a job. But notice what this "vanilla" recommender system of the sort computer scientists build regularly ignores: things like behavioral biases, market congestion, and moral hazard (Bied et al. 2023; Behaghel et al. 2024). That is, the vanilla recommender system from computer science ignores behavioral economics (and the rest of economics as well, for that matter), plus the joint design of policy and algorithm. Algorithm design is not just a computational problem; it's also an economics problem.

There is, of course, nothing about this application that is limited to unemployed workers. We see exactly the same set of issues arising in a wide range of economic applications including credit markets, health insurance, school choice, housing markets, etc. The centrality of algorithms to economics must grow over time as the centrality of algorithms to the economy and economic decisions grows over time. This is what ultimately distinguishes economics in the age of algorithms from economics in the age of electricity.

V. History Is Not Over

My essay began with a new data tool that came out of computer science: supervised machine learning. Of course, while supervised learning represents one initial major algorithmic advance, it is surely not the last. We have already seen another—the rise of large language models (LLMs). It is, therefore, not sufficient to simply have a framework for supervised learning. We need a way to think about what frameworks for *new* algorithms should look like.

To see what that might look like, it's worth looking at what we did for supervised learning. Casting them as $\hat{Y}$ tools provided two benefits. First, it allowed us to do econometrics. In this case, one could abstract from the specific details of any particular function class (LASSO versus support vector machines versus neural networks, etc.). It also made clear what econometric guarantees we needed and what assumptions we need to get those guarantees. For example, we can get bounds on the prediction loss of a learned predictor: If we're willing to assume that any new data we obtain are drawn from the same data-generating process as a holdout dataset, then our measure of the algorithm's performance in the holdout test set is informative about out-of-sample performance. We can think of that result as a "contract" for the researcher—specifying assumptions the researcher must defend in order to benefit from the guarantees provided by the result. Second, seeing that these tools were good at $\hat{Y}$ problems helped us identify new economic uses these tools can excel at.

That two-part approach can also be used going forward. The economics of algorithms is not a one-off narrow econometric activity but rather a perpetual one that starts with conceptualizing each new development's econometric properties and then connects those to all the ways in which the new tool might change the study of economic decisions or efforts to improve those decisions. To illustrate the general approach, in what follows I consider: What would this look like for LLMs?

A. *An Econometric Framework for LLMs*

As we note in Ludwig, Mullainathan, and Rambachan (2025), no econometric contracts of the sort we have for supervised learning currently exist for LLMs. The reason is that developing such contracts is particularly challenging for LLMs, partly because precisely modeling the inner workings of LLMs is difficult. LLMs are a diverse, dynamic set of extraordinarily complex machine learning models involving many layers of interactivity and billions of parameters. Their training datasets and architectures (among many other details) are often intentionally obscured because LLMs are proprietary commercial products. To further complicate matters, modeling the outputs of LLMs has proven to be equally difficult. Computer scientists struggle to characterize the brittleness of LLMs that lead their outputs to accomplish remarkable feats in some tasks (like acing the math SAT) yet produce bizarre failures in others (like hallucinating "facts" that aren't facts or failing at even very basic logical inferences like "A is B" implies that "B is A").

As with supervised learning, we require an econometric framework that is not specific to any particular LLM but rather is more general, operating at a higher level of abstraction to provide contracts

for LLMs despite the complexity of these models and without ex ante assumptions about the quality of their outputs. The framework we develop in Ludwig, Mullainathan, and Rambachan (2025) considers settings in which a researcher uses an LLM to process text and produce outputs that can be related to traditional economic variables in downstream analyses. Precisely because these algorithms are opaque and varied, we treat everything about how the LLM goes from its training data to a particular text response as a black box.

Our framework helps clarify that with LLMs we actually need two separate types of contracts, depending on the specific type of application to which the LLM is being applied.

Our first result relates to prediction problems, in which the researcher predicts a linked economic variable using the associated text. For example, in asset pricing, we might want to know how well stock prices can be predicted using news headlines. We show that for LLM outputs to be validly used in prediction problems, the LLM must satisfy a single condition we refer to as "no training leakage." Suppose the researcher prompts the LLM on each collected piece of text to form predictions and evaluates the quality of the LLM's resulting predictions by calculating its sample average loss on the researcher's dataset. We show that the LLM's sample average loss only reflects its true out-of-sample predictive performance if and only if there is no overlap between the text in the LLM's training dataset and the researcher's own dataset. Intuitively, in this workflow the researcher is using the LLM as if it were some sort of prediction function from text to the economic variable and is then using their own collected dataset as if it were a test sample. This is only valid if the LLM has not been trained on the test sample.

No training leakage is often violated in naïve uses of LLMs precisely because their training datasets are both immense and intentionally obscured. But this econometric framework suggests a solution: the use of open-source time-stamped LLMs, for which it is possible to know something about the data on which the LLM was trained.

The second type of econometric result that comes out of this framework is for what we call estimation problems, in which the researcher wants to relate some economic concept (e.g., positive or negative news, hawkish attitudes, or policy topic) expressed in text in order to estimate some downstream economic parameter. For example, in studying partisanship, how does the policy topic of a congressional bill relate to the ideology of its sponsor? We assume that the researcher has specified a resource-intensive procedure for measuring the economic concept that we would be satisfied with if it could be scaled. Because it may be prohibitively costly or time-consuming, the researcher hopes to automate their existing measurement procedure and instead use an LLM to cheaply produce these labels. The challenge here is that LLMs are brittle; their performance across tasks is, as noted above, variable and hard to predict. So how can we use LLM outputs in estimation problems knowing that they are potentially imperfect substitutes for the measurements they seek to replace?

Our econometric framework points to a constructive solution to this problem by borrowing an idea long known in economics: Collect a small sample of benchmark data and use that data to empirically model the LLM's errors. While this is well-known in labor economics (e.g., Bound and Krueger 1991; Bound et al. 1994; Bound, Brown, and Mathiowetz 2001) and well-studied in econometrics (e.g., Lee and Sepanski 1995; Chen, Hong, and Tamer 2005; Schennach 2016), it has only recently been revived in machine learning, such as Wang, McCormick, and Leek (2020); Angelopoulos et al. (2023); Wei and Malik (2022); Egami et al. (2022); and Battaglia et al. (2024), among others.

This econometric framework gives us the equivalent of a $\hat{Y}$ and $\hat{\beta}$ abstraction for LLMs similar to what we have for supervised learning and also provides econometric contracts for use in economic research (this assumption is required for this sort of performance guarantee). The key part of the contract is to now treat language as objects for analysis the same way we normally do with numbers in a table.

B. *For What Decision Applications Is This Useful?*

As with supervised learning, to use LLMs in economics, the next step after developing an econometric framework is to consider what sorts of economic decisions this new tool is most useful for understanding and potentially improving. With supervised learning, the key was to identify decisions

that hinge on a prediction of some $\hat{Y}$. With LLMs, we're now dealing with language, not numbers, so the goal is to identify economic decisions that hinge on language. LLMs are so new that we don't have concrete examples that have already been worked through end to end, so, in a forward-looking way, I will talk through some hypothetical examples intended to highlight the potential of LLMs for economic decision-making.

There is no shortage of research from behavioral economics highlighting how and why so many decisions are so hard for so many consumers. People are faced with a set of choices about what car to buy, what health insurance plan to choose at work or through some government program, what sorts of investment vehicles should hold their retirement savings, etc. Consumers might overattend to certain product attributes, such as those more easily quantified (Chang et al. 2024). They might ignore costly add-on prices that companies try to shroud from view (Gabaix and Laibson 2006). They might get overwhelmed by the sheer volume and complexity of the choice set (Kling et al. 2012). I could go on and on.

What would a potential "fix" look like? It would require explaining to consumers what in the world is going on. That would require a data tool that could make sense of the high-dimensional text description of the choice options and the customer's circumstances, identify the most likely sources of misunderstanding or error on the part of the customer, and provide some sort of comprehensible explanation of the potential upsides and pitfalls of each option. Solving these behavioral biases in practice would, in other words, inevitably require the capacity to work with language as both inputs and outputs—that is, what an LLM is designed for.

Let me offer one other illustrative example. Behavioral biases are central not only for consumer decisions but for social policy as well. For example, in Heller et al. (2017), we worked with a Chicago-area nonprofit to study an intervention called Becoming a Man (BAM) that helps de-bias young men in high-stakes situations to help prevent social harms. The result of this and similar programs can be large changes in outcomes like violence involvement, high school graduation, etc. (Bhatt et al. 2024; Abdul-Razzak and Hallberg 2024; Ludwig 2025). The challenge with BAM, as with almost everything else in social policy, is scale-up (Davis et al. 2017; Bhatt et al. 2021; List 2022).

LLMs may provide one solution to this scale-up challenge by helping automate at least parts of the program delivery. With social programs like this, we already have a curriculum that identifies key behavioral biases and ways of explaining how to avoid them. Different program participants may respond in a wide variety of different ways, depending on their circumstances, which then requires the program provider to in turn know what the best response is in each contingency. This seems to be one of the key delivery skills that is hard to scale. But this is also exactly the sort of text input–text output problem that LLMs are designed to solve. And in fact some evidence that this type of algorithmic intervention may be possible comes from a study that uses this type of technology to deliver BAM-like content to a different group of people who can also sometimes struggle with high-stakes decisions on Chicago's South and West Sides—Chicago police officers (Dube, MacArthur, and Shah 2025).

Notice also how the economic research enterprise changes once the curriculum goes from a paper document to a text object delivered by an algorithm. This shift makes the curriculum not just automatable in its delivery but *manipulable* as well. The nature of hypothesis testing and experimentation can now fundamentally change. Our traditional randomized controlled trials involve handing a set of paper curricula to a set of human program providers, having them improvise and deliver that curriculum in all manner of different ways that are nearly impossible to fully measure, then estimating the average effect of an intervention that can't really be adequately characterized. But with algorithmic delivery, we can not only perfectly characterize the intervention as it is delivered at large scale, we can (subject to sample size constraints) experimentally manipulate each of the intervention's component parts to optimize impacts.

The larger lesson here is not specific to LLMs but rather is about a general playbook for economists to follow as new artificial intelligence technologies come online in the future. The first step must be to develop an econometric framework that provides researchers with a contract—what sort of assumptions provide what sorts of guarantees. Since computer science is often at least as much about engineering as about science, this sort of econometric work will inevitably require economists to be

involved. And once that framework is developed, the next step, even more reliant on economists for obvious reasons, is to thoughtfully consider the set of economic decisions out in the world where our study of or efforts to improve them might benefit.

VI. Conclusion

A city can grow in one of two ways. The first is upward. The average height of a Manhattan skyscraper in 1900 was just 60 feet. Today, the average building there stands nearly 1,000 feet.[11] The second way cities grow is outward. While New York City is compressed into something like 300 square miles total (not counting waterways), Los Angeles sprawls across an area that's 50 percent bigger, and Houston is more than twice as large.

Scientific disciplines grow in the same two ways. Physics has largely gone upward. The methods, tools, and theories have radically changed in astronomy, but the objects of scientific interest—the vast universe every human ponders looking up at the night sky—is the same as in Ptolemy's time. Psychology, in contrast, has largely grown outward. Looking at psychology today, Freud or Jung would not quite know what to make of research on biases and heuristics, or fMRI studies mapping cognitive functions in the brain, not just because they are methodologically novel but because the very questions seem so alien.

Economics is, in my judgment, one of the few disciplines that grows both up and out. We clearly grow upward. Smith, Marshall, and Keynes would recognize and appreciate the advances we've made in understanding core issues like trade and economic growth. But we've clearly also grown outward; economists from a century ago would feel lost in much of the new territory we occupy, including on topics like game theory, the economics of the family, education, crime, discrimination, "nudges," or randomized controlled trials.

Algorithms are a powerful force for outward expansion. Like game theory and information economics, they change how we model human decision-making. Like the "credibility revolution" (Angrist and Pischke 2010) and randomized controlled trials, they change how we measure and study decisions and their consequences out in the world. Like market design and behavioral economics, they provide us with new tools for changing (not just studying) the world. Perhaps what is unique about the expansion induced by algorithms is that they are at once tools for advancement on *each one* of these margins—modeling decisions, understanding the world, and even changing the world as well.

Outward growth has a certain predictable rhythm. In the short run, every new expansion is met with the same question: "Is this economics?" The questioner is usually not even asking a question, but implicitly asserting an answer ("No"). That is because new tools are typically judged through the lens of old problems, and that is not where the most powerful new tools shine.

In the long run, for each of the major advances in economics, such concerns eventually yielded. Because the new problems proved to be every bit as interesting and important as the old problems, the territory of economics expanded. I am confident that this is what will happen with algorithms as well. It is only a matter of time until the question "Is this economics?" is eventually replaced with a statement: "This is economics."

REFERENCES

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. 2016. "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care." *American Economic Review* 106 (12): 3730–64.

Abdul-Razzak, Nour, and Kelly Hallberg. 2024. "Unpacking the Impacts of a Youth Behavioral Health Intervention: Experimental Evidence from Chicago." Annenberg Institute at Brown University EdWorkingPaper 24-1053.

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. 2022. "Machine Learning and Phone Data Can Improve Targeting of Humanitarian Aid." *Nature* 603 (7903): 864–70.

[11] https://secretnyc.co/tallest-skyline-nyc/

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-Powered Inference." *Science* 382 (6671): 669–74.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Identification and Estimation of Local Average Treatment Effects." NBER Working Paper 0118.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–932.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–60.

Autor, David. 2014. "Polanyi's Paradox and the Shape of Employment Growth." NBER Working Paper 20485.

Bald, Anthony, Eric Chyn, Justine S. Hastings, and Margarita Machelett. 2019. "The Causal Impact of Removing Children from Abusive and Neglectful Homes." NBER Working Paper 25419.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher. 2024. "Inference for Regression with Variables Generated from Unstructured Data." Preprint, arXiv. https://doi.org/10.48550/arXiv.2402.15585.

Battaglini, Marco, Luigi Guiso, Chiara Lacava, Douglas L. Miller, and Eleonora Patacchini. 2024. "Refining Public Policies with Machine Learning: The Case of Tax Auditing." *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2024.105847.

Behaghel, Luc, Sofia Dromundo, Marc Gurgand, Yagan Hazard, and Thomas Zuber. 2024. "The Potential of Recommender Systems for Directing Job Search: A Large-Scale Experiment." IZA Discussion Paper 16781.

Bergman, Peter, Elizabeth Kopko, and Julio E. Rodriguez. 2021. "A Seven-College Experiment Using Algorithms to Track Students: Impacts and Implications for Equity and Fairness." NBER Working Paper 28948.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–75.

Bhatt, Monica P., Jonathan Guryan, Salman A. Khan, Michael LaForest-Tucker, and Bhavya Mishra. 2024. "Can Technology Facilitate Scale? Evidence from a Randomized Evaluation of High Dosage Tutoring." NBER Working Paper 32510.

Bhatt, Monica P., Jonathan Guryan, Jens Ludwig, and Anuj K. Shah. 2021. "Scope Challenges to Social Impact." NBER Working Paper 28406.

Bied, Guillaume, Solal Nathan, Elia Perennes, Morgane Hoffmann, Philippe Caillou, Bruno Crépon, Christophe Gaillac, and Michèle Sebag. 2023. "Toward Job Recommendation for All." In *Proceedings of the Thirty-Second International Joint Conferences on Artificial Intelligence*, edited by Edith Elkind, 5906–14. International Joint Conferences on Artificial Intelligence Organization.

Blattner, Laura, and Scott Nelson. 2021. "How Costly Is Noise? Data and Disparities in Consumer Credit." Preprint, arXiv. https://doi.org/10.48550/arXiv.2105.07554.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–76.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics* 127 (3): 1243–85.

Bound, John, Charles Brown, Greg J. Duncan, and Willard L. Rodgers. 1994. "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data." *Journal of Labor Economics* 12 (3): 345–68.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, Vol. 5, edited by James J. Heckman and Edward Leamer, 3705–843. Elsevier.

Bound, John, and Alan B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9 (1): 1–24.

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33 (2): 220–39.

Butler, Keith T., Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. "Machine Learning for Molecular and Materials Science." *Nature* 559 (7715): 547–55.

Callen, Michael, Miguel Fajardo-Steinhäuser, Michael G. Findley, and Tarek Ghani. 2024. "Can Digital Aid Deliver during Humanitarian Crises?" Preprint, arXiv. https://doi.org/10.48550/arXiv.2313.13432.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124–27.

Chang, Linda W., Erika L. Kirgios, Sendhil Mullainathan, and Katherine L. Milkman. 2024. "Does Counting Change What Counts? Quantification Fixation Biases Decision-Making." *Proceedings of the National Academy of Sciences* 121 (46): e2400215121.

Chen, Xiaohong, Han Hong, and Elie Tamer. 2005. "Measurement Error Models with Auxiliary Data." *Review of Economic Studies* 72 (2): 343–66.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–68.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81, edited by Sorelle A. Friedler and Christo Wilson, 134–48. PMLR.

Coleman, James S. 1968. "Equality of Educational Opportunity." *Equity and Excellence in Education* 6 (5): 19–28.

Currie, Janet, and W. Bentley MacLeod. 2017. "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians." *Journal of Labor Economics* 35 (1): 1–43.

Davis, Jonathan M. V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. "The Economics of Scale-Up." NBER Working Paper 23925.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical versus Actuarial Judgment." *Science* 243 (4899): 1668–74.

Daysal, N. Meltem, Sendhil Mullainathan, Ziad Obermeyer, Suproteem K. Sarkar, and Mircea Trandafir. 2022. "An Economic Approach to Machine Learning in Health Policy." Center for Economic Behavior and Inequality Working Paper 24/22.

Dell, Melissa, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. "American Stories: A Large-Scale Structured Text Dataset of Historical US Newspapers." In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, 80744–72. Curran Associates.

Dietvorst, Berkeley J., and Soaham Bharti. 2020. "People Reject Algorithms in Uncertain Decision Domains because They Have Diminishing Sensitivity to Forecasting Error." *Psychological Science* 31 (10): 1302–14.

Dietvorst, Berkeley J., Joseph Simmons, and Cade Massey. 2014. "Understanding Algorithm Aversion: Forecasters Erroneously Avoid Algorithms after Seeing Them Err." Paper presented at the Academy of Management Annual Meeting, Philadelphia, August 5.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144 (1): 114–26.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them." *Management Science* 64 (3): 1155–70.

Doyle, Joseph J., Jr. 2007. "Child Protection and Child Outcomes: Measuring the Effects of Foster Care." *American Economic Review* 97 (5): 1583–610.

Doyle, Joseph J., Jr. 2008. "Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care." *Journal of Political Economy* 116 (4): 746–70.

Dube, Oeindrila, Sandy Jo MacArthur, and Anuj K. Shah. 2025. "A Cognitive View of Policing." *Quarterly Journal of Economics* 140 (1): 745–91.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. "How to Make Causal Inferences Using Texts." *Science Advances* 8 (42): eabg2652.

Elzayn, Hadi, Evelyn Smith, Thomas Hertz, Cameron Guage, Arun Ramesh, Robin Fisher, Daniel E. Ho, and Jacob Goldin. 2025. "Measuring and Mitigating Racial Disparities in Tax Audits." *Quarterly Journal of Economics* 140 (1): 113–63.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty." *Quarterly Journal of Economics* 138 (4): 2021–67.

Erel, Isil, Léa H. Stern, Chenhao Tan, and Michael S. Weisbach. 2021. "Selecting Directors Using Machine Learning." *Review of Financial Studies* 34 (7): 3226–64.

Eren, Ozkan, and Naci Mocan. 2018. "Emotional Judges and Unlucky Juveniles." *American Economic Journal: Applied Economics* 10 (3): 171–205.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy* 130 (4): 956–90.

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

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.

Grimon, Marie-Pascale, and Christopher Mills. 2025. "Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection." Preprint, arXiv. https://doi.org/10.48550/arXiv.2502.08501.

Hanushek, Eric A. 1999. "The Evidence on Class Size." In *Earning and Learning: How Schools Matter*, edited by Susan E. Mayer and Paul E. Peterson, 131–68. Brookings Institution Press.

Harless, David W., and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica* 62 (6): 1251–89.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. 2017. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *Quarterly Journal of Economics* 132 (1): 1–54.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2020. "A Unified Welfare Analysis of Government Policies." *Quarterly Journal of Economics* 135 (3): 1209–318.

Jabri, Ranae. 2021. "Algorithmic Policing." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4275083.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2015. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." NBER Working Paper 20847.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94.

Jencks, Christopher, Marshall Smith, Henry Acland, Mary Jo Bane, David Cohrn, Herbert Gintis, Barbara Heyns, and Stephan Michelson. 1972. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. Basic Books.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89.

Jussupow, Ekaterina, Izak Benbasat, and Armin Heinzl. 2020. "Why Are We Averse towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion." Paper presented at the European Conference on Information Systems, Marrakech.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Macmillan.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Hachette Book Group.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119.

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel. 2012. "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127 (1): 199–235.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Krueger, Alan B. 2003. "Economic Considerations and Class Size." *Economic Journal* 113 (485): F34–63.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–65.

Le Barbanchon, Thomas, Diego Ubfal, and Federico Araya. 2023. "The Effects of Working While in School: Evidence from Employment Lotteries." *American Economic Journal: Applied Economics* 15 (1): 383–410.

Lee, Lung-Fei, and Jungsywan H. Sepanski. 1995. "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data." *Journal of the American Statistical Association* 90 (429): 130–40.

Li, Danielle, Lindsey R. Raymond, and Peter Bergman. 2024. "Hiring as Exploration." Unpublished.

List, John A. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Crown Currency.

Ludwig, Jens. 2025. *Unforgiving Places: The Unexpected Origins of American Gun Violence*. University of Chicago Press.

Ludwig, Jens, and Sendhil Mullainathan. 2021. "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System." *Journal of Economic Perspectives* 35 (4): 71–96.

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation." *Quarterly Journal of Economics* 139 (2): 751–827.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2024. "The Unreasonable Effectiveness of Algorithms." *AEA Papers and Proceedings* 114: 623–27.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2025. "Large Language Models: An Applied Econometric Framework." NBER Working Paper 33344.

Meehl, Paul E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.

Merchant, Amil, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. "Scaling Deep Learning for Materials Discovery." *Nature* 624 (7990): 80–85.

Mohler, George O., Martin B. Short, Sean Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi, and P. Jeffrey Brantingham. 2015. "Randomized Controlled Field Trials of Predictive Policing." *Journal of the American Statistical Association* 110 (512): 1399–411.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics* 137 (2): 679–727.

Mullainathan, Sendhil, and Ashesh Rambachan. 2024. "From Predictive Algorithms to Automatic Generation of Anomalies." NBER Working Paper 32422.

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

Nisbett, Richard E., and Eugene Borgida. 1975. "Attribution and the Psychology of Prediction." *Journal of Personality and Social Psychology* 32 (5): 932–43.

Ohlin, Lloyd E., and Otis Dudley Duncan. 1949. "The Efficiency of Prediction in Criminology." *American Journal of Sociology* 54 (5): 441–52.

Oprea, Ryan. 2022. "Simplicity Equivalents." Unpublished.

Puri, Indira. 2025. "Simplicity and Risk." *Journal of Finance* 80 (2): 1029–80.

Rambachan, Ashesh. 2024. "Identifying Prediction Mistakes in Observational Data." *Quarterly Journal of Economics* 139 (3): 1665–711.

Rambachan, Ashesh, Amanda Coston, and Edward Kennedy. 2024. "Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding." Preprint, arXiv. https://doi.org/10.48550/arXiv.2212.09844.

Romera-Paredes, Bernardino, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, et al. 2024. "Mathematical Discoveries from Program Search with Large Language Models." *Nature* 625 (7995): 468–75.

Schennach, Susanne M. 2016. "Recent Advances in the Measurement Error Literature." *Annual Review of Economics* 8: 341–77.

Sunstein, Cass R. 2021. "Governing by Algorithm? No Noise and (Potentially) Less Bias." *Duke Law Journal* 71 (6): 1175.

Tetlock, Philip E., and Dan Gardner. 2016. *Superforecasting: The Art and Science of Prediction*. Random House.

Thaler, Richard H. 2016. "Behavioral Economics: Past, Present, and Future." *American Economic Review* 106 (7): 1577–600.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.

Tversky, Amos, and Daniel Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76 (2): 105–10.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5 (2): 207–32.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.

Wang, Siruo, Tyler H. McCormick, and Jeffrey T. Leek. 2020. "Methods for Correcting Inference Based on Outcomes Predicted by Machine Learning." *Proceedings of the National Academy of Sciences* 117 (48): 30266–75.

Wei, Max, and Nikhil Malik. 2022. "Unstructured Data, Econometric Models, and Estimation Bias." USC Marshall School of Business Research Paper Sponsored by iORB.