WORKING PAPER · NO. 2024-11

The Unreasonable Effectiveness of Algorithms

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan

FEBRUARY 2024

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

THE UNREASONABLE EFFECTIVENESS OF ALGORITHMS

Jens Ludwig
Sendhil Mullainathan
Ashesh Rambachan

This is a longer version of a paper forthcoming in the 2024 American Economic Association Papers & Proceedings. Thanks to Alejandro Roemer and Josh Schwartzstein for invaluable assistance; to Nathan Hendren, Paul Goldsmith-Pinkham, Greg Stoddard, Crystal Yang and participants in our AEA session for valuable comments; and to the Center for Applied AI at the University of Chicago and the University of Chicago Crime Lab for financial assistance. Any errors and all opinions are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Unreasonable Effectiveness of Algorithms
Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan
February 2024
JEL No. C45,C54,D61,H40,I00,K00

ABSTRACT

We calculate the social return on algorithmic interventions (specifically their Marginal Value of Public Funds) across multiple domains of interest to economists—regulation, criminal justice, medicine, and education. Though these algorithms are different, the results are similar and striking. Each one has an MVPF of infinity: not only does it produce large benefits, it provides a "free lunch." We do not take these numbers to mean these interventions ought to be necessarily scaled, but rather that much more R&D should be devoted to developing and carefully evaluating algorithmic solutions to policy problems.

Jens Ludwig
Harris School of Public Policy
University of Chicago
1307 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

Sendhil Mullainathan
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
sendhil.mullainathan@gmail.com

Ashesh Rambachan
Department of Economics
MIT
Cambridge, MA
ashesh.a.rambachan@gmail.com

I. Introduction

Are algorithms getting too much attention within economics? Bubbles arise when valuation exceeds fundamentals, when enthusiasm for *what might be* overtakes *what actually is.* Especially with the hype around large language models, are we gripped by, as Keynes might put it, "algorithmic spirits?"

To answer this question we look, just as we would with a stock, at the *fundamentals*: what tangible value do algorithms create in addressing economic issues? To answer this question we focus on their effectiveness in addressing public policy problems—the sort of algorithm that is migrating from the online world to real-world domains of traditional interest to economists. Public finance provides a direct way to measure these fundamentals: the ratio of net benefit to society by the net cost to the government. Formalized as Marginal Value of Public Funds (MVPF), these calculations have proven helpful not just in guiding policy but also in guiding policy R&D (Hendren and Sprung-Keyser 2020, 2022).

For example, encouraging results from a small-scale pilot study of class size reduction in Indiana led to the large-scale randomized controlled trial (RCT) of Tennessee STAR (Mosteller, 1995). Enthusiasm about the Perry Preschool pilot study helped motivate Congress in 1998 to support a large-scale national RCT of Head Start.[1] For the early stages of the 'R&D pipeline'—interpreting pilots to decide what is worth a large-scale RCT, for example—these calculations need only direct us to policies that have high *potential* upside, worthy of further exploration.

In this paper, we consider MVPFs for algorithmic interventions to similarly provide guidance on how much effort we should be putting into exploring algorithmic solutions to policy

[1] https://www.brookings.edu/articles/does-head-start-work-the-debate-over-the-head-start-impact-study-explained/

problems.[2] For example, Bergman et al. (2023) study a potential source of inefficiency in college course selection: if a course is too hard the student might drop out, but if it's too easy it's a waste of the student's time and money. Using an algorithm to predict course success improves course selection, so much so that the MVPF of this intervention is infinite—the policy not only generates large benefits, it is also a "free lunch" to the government.

We find similarly high MVPFs in all the other cases we study: algorithms used to make pretrial release decisions within the justice system, refer medical patients for testing, and guide workplace safety inspections. All these algorithms produce infinite MVPFs and are also a free lunch. Compared to other policies, these MVPF values all fall in the top 15% of the Policy Impacts MVPF library.[3]

The cost-effectiveness numbers for algorithms are not just remarkably large; they might even seem *unreasonably* large. Two reasons suggest that such large MVPF values are plausible, however.

The first reason stems from the logic of ranking problems, which are at the heart of so many economically important decisions (whom to hire, admit, give a loan, detain awaiting trial, etc.). The usual logic of policy interventions assumes there is some downward-sloping marginal benefit schedule and that the government has already capitalized on many of the highest-benefit cases, so expansions of the policy serve marginal cases with (relatively speaking) lower benefits,

[2] Asking about the cost-effectiveness of algorithms as a category is, in one sense, about as sensible as asking about the cost-effectiveness of "drugs." The answer of course depends on which drug, to what purpose. It is well understood by now that poorly designed algorithms used for wrong purposes can produce negative and sometimes disastrous outcomes. Here we focus on the set of algorithms that are carefully constructed, properly aligned with the objectives of policymakers, and ideally also rigorously evaluated.

[3] As of this writing there are 130 policies for which Policy Impacts reports MVPF values in the library, of which 19 have estimated MVPF values of infinity.

as shown by the Harberger triangle in the left-hand panel of Figure 1.[4] That logic assumes the government has properly rank-ordered cases by marginal benefit. But, as a large body of research now shows, government agencies and decision-makers regularly misrank. The algorithm, by improving rank-ordering of cases by marginal benefit, yields a steeper social returns schedule and a sizable reduction in deadweight loss as shown on the right in Figure 1 between the flatter and steeper schedule.

The second key reason algorithms can yield such high MVPFs is because they operate at scale. One of the key challenges with traditional policies is the scale-up problem. As the distinguished sociologist Peter Rossi described it: "Given that a treatment is effective in a pilot test does not mean that when turned over to YOAA [Your Ordinary American Agency], effectiveness can be maintained … There is a big difference between running a program on a small scale with highly skilled and very devoted personnel and running a program with the lesser skilled and less devoted personnel that YOAA ordinarily has at its disposal" (Rossi, 1987). In contrast, algorithms are software and can be run over and over again at low marginal cost without loss of fidelity. There is not the same problem of diminishing marginal returns with scale that we typically face with so many "traditional" policies (Davis et al., 2017; List, 2022).

Of course, these encouraging results are not without caveats. For example, we do not know how decision-makers will respond to algorithms in a given context. But recall our goal. We are not arguing that the government should start scaling algorithms with high MVPFs. We are arguing instead that these algorithmic policies are worth further exploration and R&D, since these are at least as promising as other policies economists work on. The caveats now serve a

[4] Figure 1 is inspired by the frameworks in Baicker et al. (2015) and Handel and Schwartzstein (2018). It can be microfounded by a simple model in which a policymaker selects between two imperfect predictors of true value, a noisy human versus a more precise algorithm. We are grateful for Josh Schwartzstein for helpful discussions.

useful purpose: they point to what should be carefully tracked and measured as part of the next stage of R&D. So, is there *too much* attention being paid to algorithms? These calculations suggest that—at least within policy applications—algorithms are receiving *too little* attention.

We begin in Section II with a case study of an algorithm applied to pretrial release decisions within the criminal justice system; when we calculate the different components of the MVPF we get an estimated value of infinity. We then show in Section III that the remarkable MVPF of the New York City pre-trial algorithm is not a "unicorn," by considering a number of additional algorithmic policy interventions as well. Each of their MVPF values is also infinity. These figures compare favorably to the catalog of MVPF values for traditional policies that has been assembled by Policy Impacts.[5] However, as we note in section IV, there remain a number of important conceptual and econometric problems that remain to be solved for both the algorithmic R&D pipeline and efforts to take algorithms to scale as policies. Section V concludes.

II. Pretrial Release

We begin with a detailed case study of an algorithm applied within the criminal justice system, to judges' decisions about whom to release from jail awaiting trial. Police in the US make something like 10 million arrests each year. After a defendant is arrested they go before a judge within 48 hours, whose job it is to decide where they await trial—at home or in jail.[6] By law, that decision is supposed to be based on the judge's prediction of the defendant's risk of

[5] https://www.policyimpacts.org/

[6] Technically the judge typically has multiple options: release without conditions; release with conditions; or set bail. In practice, most people who have bail set on them will spend at least some days in jail (and many will spend their whole pretrial term in jail) even if the bail amount is only a few hundred dollars. For the sake of simplicity, we discuss this as a release-detain decision.

skipping court or re-offending, based on factors like the charge for which the defendant was arrested or their prior criminal record (Dobbie and Yang, 2021).

Kleinberg et al. (2018) note that these predictions could in principle be made by an algorithm instead. This application has all the key ingredients that make it suitable for construction of an algorithmic decision aid: a large number of cases; a great deal of information available about each case; and a socially important decision that hinges on a predictive inference.

This application also nicely illustrates both the benefits of re-ranking infra-marginal cases. Using data from New York City, Kleinberg et al. show that though a very large share of defendants present close to zero risk, judges send about 10% of these cases to jail.[7] Conversely, there are some cases with a predictable ex ante risk of failure as high as four in five, yet judges let nearly 50% of these cases go. Of course there are ongoing debates about whether anyone should be detained pretrial, with many calling for major overhauls to the system. But with the existing system in place for the foreseeable future, re-ranking cases by more accurate risk predictions creates the potential for sizable benefits to society from both fewer jail spells and fewer crimes and failures to appear (FTA) in court.

Our goal here is to go from evidence of human mis-ranking to a calculation of the MVPF, defined by Hendren and Sprung-Keyser (2020) as the benefit to society divided by net cost to government or MVPF = $\Delta W / (\Delta E - \Delta C)$, where ΔW is the value of policy impacts on affected people (willingness to pay), ΔE is the up-front change in government expenditures (for example, to build and deploy some new algorithm) and ΔC is any savings to government spending

[7] In New York State, the law says judges are supposed to only focus on the risk of failure to appear in court (FTA) not on risk of re-arrest or violence.

achieved by the policy.[8] This calculation involves at least three sources of uncertainty that are difficult to quantify absent data from a deployed algorithm.

The first source of uncertainty comes from quantifying the benefits of the algorithm, both to the population (the public and defendants alike), ΔW, and as savings to the government, ΔC. Many papers provide proof of concept of the potential gains from an algorithm by comparing hypothetical algorithmic decisions with actual human decisions using retrospective data from past cases. But that means that outcomes measuring pre-trial failure (like re-arrest or FTA) can only be observed for defendants the judges decided to release, the so-called 'selective labels' problem (Kleinberg et al., 2018a, Rambachan et al., 2023). Even in cases where this problem can be overcome, retrospective data can't tell us anything about human compliance with any new algorithmic decision-aid. That can't be known until the algorithm is actually deployed.[9]

A second source of uncertainty comes from the cost of building the algorithm, ΔE. This also cannot be directly quantified absent an algorithm that's been built in the real world.

A third source of uncertainty comes from the fact that ultimately the government has the choice of points in the tradeoff space. For example, in the case of improved ranking of defendants for pretrial release decisions, it is possible to take the potential gains all in the form of reduced crime and FTA (holding the detention rate constant relative to status quo), or all in the form of reduced detention rates (holding crime and FTA constant), or as some combination of

[8] For a discussion of the MVPF versus other candidate social welfare criteria, see Hendren and Sprung-Keyser (2020, 2022) and Garcia and Heckman (2022).

[9] There are of course a number of other key issues that arise as well, like the challenge of knowing whether the algorithm has improved the utility of the decision-maker given that objective is usually left implicit rather than directly observed; the so-called 'omitted payoffs' problem (Kleinberg et al., 2018a, Rambachan, 2023); and the possibility that bias in the underlying criminal justice data can lead to bias in the algorithmic decision-aid itself, although how that compares to whatever bias a human would add when based with the same biased data is an application-specific empirical question (Kleinberg et al., 2018b, Obermeyer et al., 2019). We background those issues here not because they are unimportant, but rather because they are so important they warrant more extensive discussion in papers that focus explicitly on these key evaluation issues.

reduced crime and FTA and jail. This decision is relevant for the MVPF calculation because it can determine how the benefits of the policy are distributed, not just across different groups within the population but between the public versus the government budget. This also highlights that even the shift to algorithmic policy tools does not eliminate the room for (even the need for) policymakers to decide key normative policy questions.[10]

Despite these sources of uncertainty, progress in calculating MVPFs for algorithms is often still possible because even conservative estimates typically yield quite favorable figures. Kleinberg et al. (2018) present a policy simulation showing that letting the algorithm rather than the judge make release decisions could allow for up to 40% fewer pretrial detentions with no increase in pretrial failure rates (crime or FTA). Because this policy simulation holds pretrial failure rates constant, in terms of the public's willingness to pay for the algorithm, call this ΔW_P, we argue that this should be positive (or at least non-negative) for all population sub-groups even if we cannot directly quantify these values. The other population whose willingness to pay is relevant is the defendants who would have been detained absent the new algorithm and now get to go home instead. We estimate the value of freedom and higher labor market earnings together equal $3,200 per jail spell averted.[11] Note these benefits accrue disproportionately to the socially disadvantaged groups currently greatly over-represented in New York jails: Black defendants (57% of jail inmates; Kleinberg et al. 2018) and Hispanic defendants (32%). If we conservatively assume judges using the algorithm yield one-quarter the benefits of the algorithmic decisions

[10] Thanks to our discussant, Paul Goldsmith-Pinkham, for highlighting this point.

[11] Abrams and Rohlfs (2011) estimate the typical defendant would pay $1,000 in 2003 dollars to avoid a jail spell of 90 days. Adjusting for inflation to bring this up to 2023 dollars,assuming willingness to pay scales linearly in the length of the jail spell, and prorating this to match the 115 day average spell in Rikers Island, we estimate $2,200 in current dollars per NYC jail spell averted. The causal effect of a jail spell on earnings from Dobbie et al. (2018) are somewhat imprecisely estimated but the point estimates, taken at face value, imply that avoiding a jail spell leads to about $1,000 higher income in the formal labor market by three or four years afterwards, a sizable change relative to a control mean of only around $6,000. (Note that accounting for the additional tax revenue from this increase in earnings would serve to further increase the value of government savings, ΔC.)

themselves,[12] with 20,000 arraignments per year and assuming a 50% release rate for the type of cases still eligible for bail hearings in New York this implies 1,000 fewer detentions and a numerator for the MVPF of (ΔW_P + $3.2 million).

Our best estimate for ΔC from 1,000 fewer detentions is savings to the government on the order of $34.5 million per year, calculated as follows:

- The average jail spell on Rikers Island has been reported to be 115 days,[13] while the average cost per person per day in jail has been reported to be on the order of $1,500.[14]
- Of course the marginal cost will be far below the average cost. As the Rikers Island jail population has plummeted over time, for example, the size of the jail staff has not declined commensurately; in fact, staffing has hardly changed. We use an estimate from the Vera Institute that marginal costs in the criminal justice system in general may be on the order of 20% of average costs.[15]
- Together these figures imply a reduction in government spending for each averted jail spell that equals $34,500. (Note this is a lower-bound estimate in that it excludes the additional tax revenues derived from increased earnings from those who would have been detained absent the algorithm, but now are freed instead).

Putting this together implies the MVPF will be infinite so long as the costs of the algorithm, ΔE, are less than the government savings, ΔC=$34.5 million:

$$MVPF = (\Delta W_P + \$3.2m) / (\Delta E - \$34.5m)$$

[12] Albright (2023) studies how judges respond to an algorithmic decision aid in the Kentucky pretrial release system, finding that algorithmic recommendations raise release rates by 15 percentage points.
[13] https://gothamist.com/news/detainees-spend-an-average-of-115-days-at-rikers-4-times-the-national-average
[14] https://comptroller.nyc.gov/newsroom/comptroller-stringer-cost-of-incarceration-per-person-in-new-york-city-skyrockets-to-all-time-high-2/
[15] https://www.vera.org/downloads/publications/marginal-costs-guide-fact-sheet.pdf

While these results come from a proof-of-concept policy simulation from Kleinberg et al. (2018), we can validate some of the key parameters in this case because there is a real-world instantiation of this algorithm that was actually deployed. Specifically, a few years ago the research center run by one of us (Ludwig), the University of Chicago Crime Lab, was asked by the Mayor's Office of Criminal Justice (MOCJ) in New York City to help update the city's algorithmic decision aid for judges making pretrial release decisions.[16]

One key source of uncertainty is what judges do with the new algorithmic tool. While the formal evaluation of the new algorithm is still in progress, some initial descriptive statistics let us ballpark the potential impact for now:

- If we take the previous algorithm, which New York City had been using since 2003, and run it through more up-to-date data, it recommends 31.7% of Black defendants for release and 41.1% of white defendants.
- The new risk tool applied to those data recommends 83.9% of Black defendants for release and 83.5% of white defendants for release (set to maximize the release rate subject to not increasing the pretrial failure rate).

Given that the share of defendants recommended for release increases by 40 or 50 percentage points with the introduction of the new algorithm, the assumption of a 10% (or 5 percentage point) change in release rates seems reasonable, if not conservative. As noted above, this implies something like 1,000 fewer jail detentions each year in New York City and an estimate of ΔC equal to $34.5 million per year.

We next consider ΔE, the direct cost of building and deploying the algorithm. Most of the relevant costs were for labor:

[16] The project team at the Crime Lab was led by Greg Stoddard, carried out in partnership with MOCJ, the Criminal Justice Agency (New York City's pre-trial organization), and Marie Van Nostrand and her team at Luminosity.

- Software engineers helped extract individual-level records from the New York State Division of Criminal Justice Services, linked to court records from the Office of Court Administration and pre-trial information from New York City's Criminal Justice Agency.
- Data scientists prepared the data for analysis, trained several candidate machine learning models, and helped identify tradeoffs between predictive accuracy and explainability.[17]
- Program managers coordinated among the different organizations involved, and helped organize meetings to solicit feedback from external stakeholders (including judges, prosecutors, defense lawyers, civil rights advocates, and police leadership).
- Data scientists and software engineers at what was then called the New York City Department of Information Technology and Telecommunications (DOITT, now called the Office of Technology and Innovation) created the 'piping' to connect the live criminal justice data feeds to the court to calculate pretrial risk scores for defendants in real time.
- Training and education was required for the practitioners involved in court proceedings.

News accounts reported the algorithm cost $2.7 million to build and deploy.[18] It is very possible these figures under-estimate costs given the intrinsic difficulties of accounting for the opportunity cost of the time of public-sector workers who were involved, including MOCJ staff time to manage the project, DOITT staff time and the various public-sector stakeholders who provided feedback and devoted time to training to use the new tool. We conservatively assume the true cost is approximately 50% higher than news reports claim, or about $4 million.

[17] Ideas42, a behavioral science non-profit co-founded by Mullainathan, helped design the user interface for the tool.

[18] https://www.wsj.com/articles/algorithm-helps-new-york-decide-who-goes-free-before-trial-11600610400

One final question for our MVPF calculation is how often the algorithm would need to be retrained to account for 'data drift' (changes in the underlying data generating process). The most conservative assumption we could make would be to assume the algorithm has to be rebuilt every year, so the city gets only one year's worth of use out of each algorithm build. To see just how conservative this is, we note that in practice the algorithm was first deployed in 2019 and as of this writing (early 2024) the same algorithm is still in use in every New York City courtroom. Using the data from the case study to verify and/or fill in parameters from the policy simulation leads to a MVPF estimate of:

$$MVPF = (\Delta W_P + \$3.2m) / (\$4m - \$34.5m) = \text{infinity}$$

Note the literal value of this social welfare calculation (infinity) is sensitive to the functional form assumption behind the MVPF formula, which is constructed to heavily weight the net cost to the government. If we used alternative social welfare measures like net benefits or a benefit-cost ratio, the values would not be infinity but the figures would nonetheless be large both absolutely and relative to other candidate policies. That is, the conclusion of very favorable cost-effectiveness does not hinge on the choice of any particular social welfare metric.

III. Additional examples

The pretrial release tool for New York City is an encouraging example and, as we will show here, not an isolated one.

A. Safety regulation

OSHA regulates workplace safety currently by targeting inspections based on the number of workplace injuries at each establishment over the past few years. Johnson et al. (2023) show that an algorithm can better predict which work sites are likely to have another injury in the future. Targeting OSHA inspections using this algorithm instead is estimated to reduce the

number of serious injuries by at least 15,934.[19] Multiplying that figure by the estimated cost per serious injury (based on the number of days of work missed) implies a benefit to workers from fewer injuries equal to at least ΔW=$844 million.

Since the average federal tax rate for Americans is 24.8%, an algorithm that prevents $844 million in lost income leads to an increase in tax revenue collection equal to ΔC= $209.3 million. Given any plausible figure for the cost of building and deploying this algorithm (denominated most likely in the single-digit millions, and certainly not more than a few tens of millions), the estimated MVPF of this algorithm is, again, infinity.

B. Health Care

Hundreds of thousands of people show up at the emergency room every year complaining of chest pain, worried they are having a heart attack. A doctor has to decide whether to refer the patient to a follow-up 'stress test' to determine whether they are actually having a heart attack. Sending a patient for testing who actually just has acid reflux (which can create similar symptoms) wastes money and the patient's time. Not sending a patient for a test who is having a heart attack can lead the patient to, in the extreme, die. The current testing decision is made by a doctor applying their best judgment to a collection of diagnostic health information (EKG results, the patient's past health history, their description of their current symptoms, etc.)

As Mullainathan and Obermeyer (2022) show, an algorithm could use those data instead to predict patients at highest risk for heart attack—and that one proof-of-concept example of such an algorithm seems to predict patient risk far more accurately than the (human) doctors do. The effects of this algorithmic re-ranking are shown in the right-hand panel of Figure 1: Doctors, by confusing low-risk patients for high-risk ones, and vice versa, essentially create a list of

[19] This estimate comes from using machine learning based predictions; using heterogeneous treatment effect estimates instead implies 16,524 serious injuries averted.

patients rank-ordered by perceived risk of heart attack (which is the marginal benefit of testing in this application) that is 'too flat' compared to the algorithmic ranking. Doctors make a second mistake, besides,which is to use too low of a threshold for the expected health gains needed to justify testing. Abstracting for now from whether the doctors would follow the algorithm's recommendations, if we combine the steeper marginal benefit schedule for testing from the algorithm's ranking with a higher marginal cost of testing threshold (consistent with the $150,000 per year of life often used in medical cost-effectiveness calculations), the implication is that we could reduce testing by 34.7% with no loss in social welfare over Medicare patients.

If we focus just on those covered by Medicare (a lower bound), the data from Mullainathan and Obermeyer (2022) imply there are 50,838 stress tests per month and 34,318 catheterizations per month. The combination of using the algorithm to re-rank patients and using a more appropriate value per life year to set the testing threshold implies 17,640 fewer stress tests per month and 11,908 fewer catheterizations per month. The Medicare fee schedule shows the cost of a stress test as $4,000 and a catheterization is $28,000. Since these are all Medicare patients whose health care costs are borne by the federal government, the new algorithmic testing rule reduces testing costs by $406 million per month or ΔC=$4.8 billion per year. The numerator ΔW is whatever patients are willing to pay to avoid the time and pain of needless tests. Even if (conservatively) the algorithm had to be rebuilt every single year, if the algorithm build cost, ΔE, is measured in the millions (or even tens or hundreds of millions), the denominator of the MVPF calculation, $\Delta E - \Delta C$, will be negative and the MVPF value of the algorithm is infinity.[20]

C. Education

[20] As with the pretrial release algorithm, we do not know how exactly doctors would respond to the introduction of such a heart attack diagnosis tool. Nonetheless, even conservative assumptions about how doctors would respond to the algorithm still imply large cost savings by this calculation.

Bergman et al. (2023) evaluate an algorithm that screens college students into college courses, and in particular the allocation of students to remedial (pre-college-level) courses in math and English versus into college-level courses. Under-placing students who are actually prepared for college-level work means they spend time and money on courses that earn them no college credit when they could instead have been working towards their degrees. Over-placing students who are not ready for college classes runs the risk of them wasting time and money on classes they fail. The current default targeting uses student scores on an academic assessment.

The study's estimates for ΔW and ΔC from algorithmic placement come from an RCT: An algorithm that considers a broader set of academic background characteristics predicts student performance more accurately than the current achievement test, increases placements into college-level classes by 2.6 percentage points in math and 13.6 pp in English (and narrows disparities across race and ethnic groups) without any decline in course pass rates. Importantly, the number of remedial credits attempted reduces by 1.1 credits and the number of college credits earned increases by 0.53 credits. The reduction in remedial credits due to algorithmic placement saves students, on average, $150, which corresponds to $145,200 savings per cohort per college. For the colleges in the experiment, the government subsidizes credit-taking, and the net change in credit-taking is estimated to produce $230 in savings per student. The authors estimate that the cost of implementing the algorithmic placement is $140 per student. The resulting MVPF calculation is then $150/($140 - $230) = infinity.

IV. Open Questions

We have intentionally been provocative in highlighting a number of remarkably effective algorithms that illustrate the enormous potential of these new technologies for social good. Of course there are many traditional policies that also have infinite MVPF values, as shown in

Figure 2 (about 15% of the 130 policies included in the Policy Impacts MVPF library).[21] While it was not particularly hard for us to come up with examples of algorithmic policies that had infinite MVPF values, without a more exhaustive effort to comprehensively calculate MVPF values for a more comprehensive set of algorithms (and traditional policies too, for that matter), it would be premature to claim that algorithms have higher MVPF values on average. Our claim here is narrower: We may be leaving many cost-effective policies on the cutting room floor by not paying more attention to algorithms as a class of policy intervention to study. We need more algorithms to enter into the R&D pipeline.

But we need another thing beyond more careful pilot studies and RCTs: We need answers to a set of fundamental economic and econometric questions that these new algorithmic tools raise. Just as the 'credibility revolution' raised new questions about heterogeneous treatment effects and compliance and local average treatment effects and equilibrium effects, the growing attention to algorithms creates new conceptual and empirical problems to solve as well. In what follows we highlight three of these key issues, the solutions to which would be of enormous value for economists as they work more in this area.

A. The algorithm's benchmark: The human

Most of the high-MVPF algorithms we examine have a key shared feature: The alternative to the algorithm is human judgment, with all its imperfections—frequent reliance on heuristics and biases (Kahneman, 2011), noise in decision-making (Kahneman et al., 2021), and out-group biases (Brewer, 1999). The result is that for a given data frame—a given set of observations and variables—the algorithm is able to extract sources of signal that humans often

[21] This was the state of the Policy Impacts MVPF library as of January 4, 2024.

cannot notice (Ludwig and Mullainathan, 2023; Mullainathan and Rambachan, 2023). Human judgment is typically such a low bar that it is easy for algorithms to soar over it.

However this need not always be the case, since, as Ludwig and Mullainathan (2021) note, in principle humans have their own source of comparative advantage over the algorithm: People often see additional information that the algorithm cannot. For example, doctors see things about patients in person that are not captured in any electronic medical record; judges hear courtroom arguments; and teachers interact with their students every day in ways that may communicate useful information not reflected in a test score or student writing sample. Understanding when people use this extra information as a source of valuable signal versus a source of unhelpful distraction is an active area of current research and something about which we desperately need to know more.[22]

B. Automation vs. Decision Aids

In many of our back-of-the-envelope MVPF calculations, we have assumed that the relevant decision is being automated (the algorithm decides). In that case the key question of social benefit hinges on the nature of the algorithm's predictive advantage over the human. But in many policy-relevant applications, the algorithm is not the decider but rather a decision-aid for some human decider. The introduction of an algorithm into a decision-making environment could in principle have no impact at all if the humans simply ignore it. Or the algorithm could even have adverse impact if humans misunderstand their comparative advantage relative to the algorithm, and for instance get distracted by irrelevant information. Existing empirical

[22] It is also the case that the dataification of so many aspects of modern life mean that simple data tools will also inevitably spread over time, so that the benchmark for the algorithm may eventually increasingly be another data tool. In the study of college student course placement by Bergman et al. (2023), the alternative to the algorithm was a very simple data rule, so the algorithm was able to still generate large relative gains. In contrast, in the study of predicting police misconduct by Stoddard et al. (2024), the algorithm has relatively more modest gains relative to a very simple count of prior misconduct events for officers.

evaluations of the implementation of different pretrial algorithms provide some evidence of each type of response.[23]

Given the diversity of findings from how humans respond to these new tools in practice, it is hard not to believe that the specific design features of these algorithms might contribute to the observed variation in algorithmic impacts across settings. Unfortunately very little is currently known on this front at present. More generally, much more needs to be known about how to help humans recognize their own and the algorithm's sources of comparative advantage in order to optimally decide when to override versus follow the algorithm's predictions (for example, Agarwal et al., 2023; Angelova, Dobbie and Yang, 2023).

C. Context-dependence

The degree to which algorithms have favorable values for the MVPF or other social welfare metrics will depend on the degree to which the algorithm's predictions generalize across settings versus are highly context-dependent. For New York City the scale of its local criminal justice system means that an algorithm with a build cost of $4 million can still easily yield a MVPF of infinity. That's less likely to be true for the much smaller jurisdictions of Scarsdale or Fort Lee or Levittown if they had to build their own algorithms from scratch, but could be true for them if a single 'small suburb risk predictor' worked well across small suburbs.

A different dimension of context besides geography is time. If the underlying data generating process differs substantially across time periods, an algorithm would need to be updated relatively more frequently, thereby increasing the build costs to the government. This

[23] For example, Stevenson (2018) shows that judges did follow the recommendation of an algorithm in Kentucky, releasing more low-risk defendants and detaining more high-risk ones, although over time judge decisions began to converge back to their previous patterns—and seem to have not responded at all to the switch in 2013 from the state's previous risk tool to a new one built by Arnold Ventures (see also Albright, 2023). Stevenson and Doleac (2022) show that a different algorithm adopted in Virginia has numerous examples of potentially unhelpful judge overrides of the new tool, and on net no detectable changes in crime rates or jailing rates.

sort of ‘data drift’ is of particular concern when adversarial actors in the world strategically respond to the algorithm (Hardt et al., 2016; Bjorkegren et al., 2023). The degree to which this is a problem will depend on the application. For example, for the New York pretrial algorithm, the main way to ‘game’ the algorithm would be for someone to strategically make their prior record or present charge less serious in nature. A synonym for people trying to dial down their crime involvement to avoid punishment is “deterrence.”

Relatively little is currently known about the degree of generalizability with algorithmic risk tools across either geography or time, although more information about that important point would be invaluable for understanding exactly the scope for social gains from this policy.[24] Relatedly, how to shore up algorithms in the presence of strategic behavior in adversarial settings is also poorly understood and is an active area of ongoing research.

V. Conclusion

If there is one lesson from the last 20 or 30 years of policy work in empirical economics, it is that there is no shortage of problems—just a shortage of solutions. Algorithms provide a whole category in which to look for new solutions.

Our claim is not that they are fool-proof. Nor that they are sure to work. Our claim is narrower: they show immense potential. And they deserve far more attention, in terms of both rigorous evaluation and careful design.

Given that problems are plentiful and solutions are scarce, there is little wonder that algorithms are receiving so much attention. They are not just particular solutions to specific

[24] An example is the prediction of police misconduct discussed in Stoddard et al. (2024), which finds that the most effective predictors of risk in the Chicago Police Department turn out to be the most effective predictors at the New York Police Department as well.

problems but represent a novel approach to solving many problems. Whether that promise bears out or not is yet to be seen. There is only one way to find out.

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.

Abrams, David S and Chris Rohlfs (2011) “Optimal bail and the value of freedom: Evidence from the Philadelphia bail experiment,” Economic Inquiry, 49(3): 750-770.

Acemoglu, Daron and Pascual Restrepo (2018) “The race between man and machine: Implications of technology for growth, factor shares, and employment.” American Economic Review. 108(6): 1488-1542.

Acemoglu, Daron, David Autor, Jonathan Hazell and Pascual Restrepo (2022) “Artificial intelligence and jobs: Evidence from online vacancies.” Journal of Labor Economics. 40(S1)

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). “Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology.” Cambridge, MA: NBER working paper 31422.

Albright, Alex (2023). “The Hidden Effects of Algorithmic Recommendations.” SSRN Working Paper.

Angelova, Victoria, Will Dobbie and Crystal Yang (2023) “Algorithmic recommendations and human discretion.” Cambridge, MA: NBER working paper 31747.

Athey, Susan, Christian Catalini and Catherine Tucker (2017) “The digital privacy paradox: Small money, small costs, small talk.” Cambridge, MA: NBER working paper 23488.

Athey, Susan, Dean Karlan, Emil Palikot and Yuan Yuan (2022) “Smiles in profiles: Improving fairness and efficiency using estimates of user preferences in online marketplaces.” Cambridge, MA: NBER working paper 30633.

Baicker, Katherine, Sendhil Mullainathan and Joshua Schwartzstein (2015) “Behavioral hazard in health insurance.” Quarterly Journal of Economics. 1623-1667.

Battaglini, Luigi Guiso, Chiara Lacava, Douglas L. Miller and Eleonora Patacchini (2022) “Refining public policies with machine learning: The case of tax auditing.” Cambridge, MA: NBER working paper 30777.

Bergman, Peter, Elizabeth Kopko and Julio E. Rodriguez (2023) “A seven-college experiment using algorithms to track students: Impacts and implications for equity and fairness.” NBER working paper 28948.

Bjorkegren, Daniel, Joshua E. Blumenstock and Samsun Knight (2023). “Training Machine Learning to Anticipate Manipulation.”

Bresnahan, Timothy F., Erik Brynjolfsson and Lorin M Hitt (2002) “Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence.” Quarterly Journal of Economics. 117(1): 339-376.

Brewer, Marilynn B. (1999) “The psychology of prejudice: Ingroup love and outgroup hate?” Journal of Social Issues. 55(3): 429-444.

Brynjolfsson, Erik, and Tom Mitchell (2017) “What can machine learning do? Workforce implications.” Science. 358(6370): 1530-1534.

Brynjolfsson, Erik, Tom Mitchell and Daniel Rock (2018) “What can machines learn and what does it mean for occupations and the economy?” AEA Papers & Proceedings. 108: 43-47.

Brynjolfsson, Erik, Danielle Li and Lindsey R. Raymond (2023) “Generative AI at work.” Cambridge, MA: NBER working paper 31161.

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: Papers & Proceedings. 106(5): 124-7.

Davis, Jonathan M.V., Jonathan Guryan, Kelly Hallberg and Jens Ludwig (2017) “The Economics of Scale-Up.” Cambridge, MA: NBER working paper 23925.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. (2018) "The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges." *American Economic Review* 108(2): 201-240.

Dobbie, Will and Crystal S Yang (2021) “The US pretrial system: Balancing individual rights and public interests.” Journal of Economic Perspectives. 35(4): 49-70.

Garcia, Jorge Luis and James J. Heckman (2022) “Three criteria for evaluating social programs.” Cambridge, MA: NBER working paper 30507.

Goldfarb, Avi and Catherine Tucker (2012) “Privacy and innovation.” Innovation Policy and the Economy, University of Chicago Press, 12(1): 65-90.

Grimon, Marie-Pascale and Christopher Mills (2023) “The impact of algorithmic tools on child protection: Evidence from a randomized controlled trial.” Notre Dame department of economics working paper.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan MV Davis, Kenneth Dodge, George Farkas, Roland G. Fryer Jr, Susan Mayer, Harold Pollack, Laurence Steinberg and Greg Stoddard (2023) “Not too late: Improving academic outcomes among adolescents.” American Economic Review. 113(3): 738-65.

Handel, Benjamin and Joshua Schwartzstein (2018) “Frictions or mental gaps: What’s behind the information we (don’t) use and when do we care?” Journal of Economic Perspectives. 32(1): 155-178.

Hardt, Moritz and Nimrod Megiddo and Christos Papadimitriou and Mary Wootters (2016) “Strategic Classification.” *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science.*

Hendren, Nathaniel and Ben Sprung-Keyser (2020) “A unified welfare analysis of government policies.” Quarterly Journal of Economics. 135(3): 129-1318.

Hendren, Nathaniel and Ben Sprung-Keyser (2022) “The case for using the MVPF in empirical welfare analysis.” Cambridge, MA: NBER working paper 30029.

Johnson, Matthew S., David I. Levine and Michael W. Toffel (2023) “Improving regulatory effectiveness through better targeting: Evidence from OSHA.” American Economic Journal: Applied Economics. 15(4): 30–67

Kahneman, Daniel (2011) Thinking, Fast and Slow.

Kahneman, Daniel, Olivier Sibony and Cass R. Sunstein (2021) Noise: A flaw in human judgment.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan and Ziad Obermeyer (2015) “Prediction policy problems.” American Economic Review: Papers & Proceedings. 105(5): 491-5.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan (2018a) “Human decisions and machine predictions.” Quarterly Journal of Economics.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan and Cass R Sunstein (2018b) “Discrimination in the age of algorithms.” Journal of Legal Analysis. 10: 113-174.

Li, Danielle, Lindsey R. Raymond and Peter Bergman (2020) “Hiring as exploration.” Cambridge, MA: NBER working paper 27736.

List, John A. (2022) The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale.

Liu, Zhuang, MIchael Sockin and Wei Ziong (2023) “Data privacy and algorithmic inequality.” Cambridge, MA: NBER working paper 31250.

Ludwig, Jens and Sendhil Mullainathan (Forthcoming) “Machine learning as a tool for hypothesis generation.” Quarterly Journal of Economics.

Mosteller, Frederick (1995) “The Tennessee Study of Class Size in the Early School Grades.” The Future of Children. 5(2): 113-127.

Mullainathan, Sendhil and Ziad Obermeyer (2022) “Diagnosing physician error: A machine learning approach to low-value care.” Quarterly Journal of Economics. 137(2): 679-727.

Mullainathan, Sendhil and Ashesh Rambachan (2023) “From Predictive Algorithms to Automatic Generation of Anomalies.” MIT working paper.

Obermeyer, Ziad, Brian Powers, Christine Vogeli and Sendhil Mullainathan (2019) “Dissecting racial bias in an algorithm used to manage the health of populations.” Science. 366(6464): 447-453.

Rambachan, Ashesh (2023) “Identifying prediction mistakes in observational data.” MIT working paper.

Rambachan, Ashesh, Amanda Coston, and Edward Kennedy (2023) “Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding.” arXiv preprint, arXiv:2212.09844.

Rossi, Peter (1987) “The iron law of evaluation and other metallic rules.” Research in Social Problems and Public Policy. 4: 3-20.

Stevenson, Megan. T. (2018) “Assessing risk assessment in action.” Minnesota Law Review. 58.

Stevenson, Megan T. and Jennifer L. Doleac (2022) “Algorithmic risk assessment in the hands of humans.” SSRN Working Paper.

Stoddard, Greg, Dylan Fitzpatrick and Jens Ludwig (2024) “Predicting police misconduct.” University of Chicago Working Paper.

Figure 1: Stylized illustration of the social welfare gains from algorithmic re-ranking of who is prioritized for services

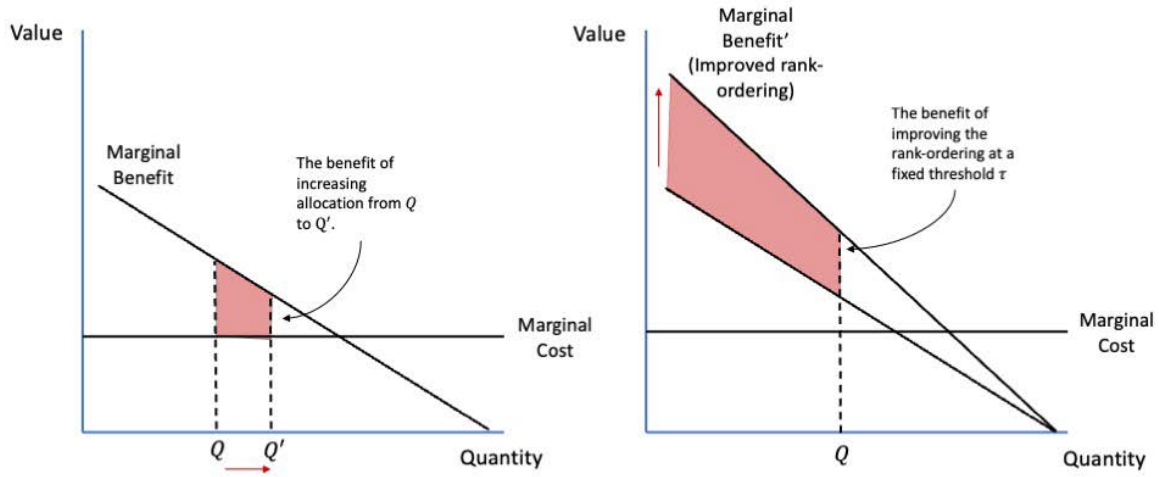
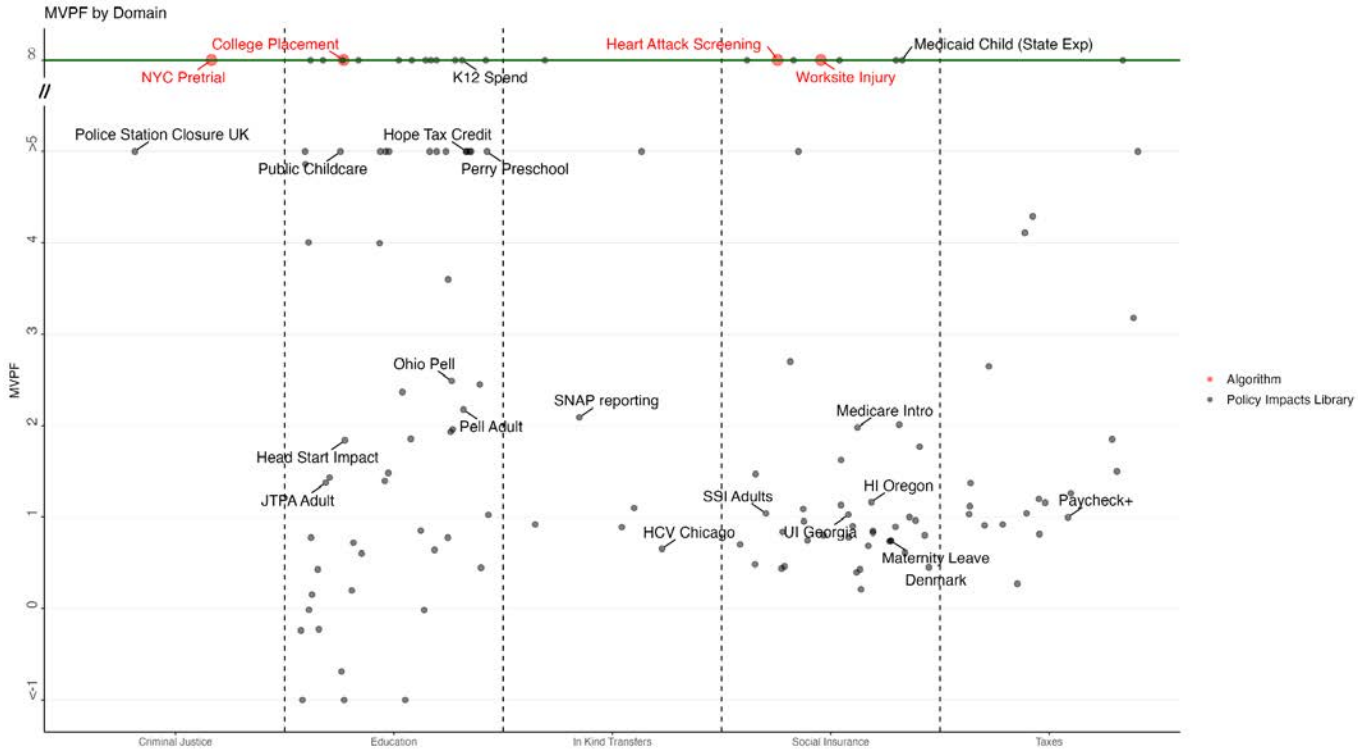


Figure 2: Comparing MVPF Values for ‘Traditional Policies’ to Those for Algorithms



Source: Authors’ calculations from MVPF calculations described in text and MVPF values taken from Policy Impacts library.