AEA Papers and Proceedings 2024, 114: 623–627
https://doi.org/10.1257/pandp.20241072

The Unreasonable Effectiveness of Algorithms[†]

By Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan*

Algorithms are receiving a lot of attention in economics, especially since the advances around large language models. Are they receiving too much attention?

To answer this question, we ask a more pragmatic one: how effective are they in addressing policy problems economists have long focused on? Familiar tools can be brought to bear now. The marginal value of public funds (MVPF) from Hendren and Sprung-Keyser (2020), defined as the ratio of net benefit to society to net cost to government, provides both a way to calculate effectiveness and, since it has been widely applied, a way to compare algorithms to more traditional public policy levers.

We calculate MVPFs for algorithms in education, criminal justice, health, and regulation. Though the particular applications are diverse and the algorithms are different, the results across them are the same and striking. Each algorithm we consider has an MVPF of infinity, meaning that each one not only produces large benefits but is also a "free lunch." Compared to other policies, these MVPF values all fall in the top 15 percent of the Policy Impacts MVPF library.

The cost-effectiveness numbers for these algorithms might seem unreasonably large, but they are plausible for two reasons. The first stems from the logic of ranking problems at the heart of many economically important decisions (whom to hire, admit, give a loan, detain awaiting trial, etc.). 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 1). 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 schedule (Figure 2). The second reason algorithms can yield high MVPFs is because they operate at scale. There is not the same diminishing marginal returns as with many traditional policies (Davis et al. 2017; List 2022).

Of course, there are caveats with these results, but we are not arguing that they are the final word. Instead, we take an iterative approach to policy. Initial research suggests where to focus further resources for investigation, and this process repeats until estimates are robust enough to scale policies. For example, pilots in early childhood education or class-size reduction motivated large-scale randomized controlled trials (RCTs).

Our takeaway from these MVPF calculations (and their caveats) is that algorithms merit further policy R&D: improved designs, careful pilots, and rigorous in situ evaluations. So, is there too much attention being paid to algorithms? These calculations suggest that, at least within policy applications, algorithms receive too little attention.

I. Pretrial Release

We start by examining an algorithm that helps judges make pretrial release decisions; more details for each of our MVPF calculations are in Ludwig, Mullainathan, and Rambachan (2024). Police in the United States make ten million arrests per year. Defendants go before a judge who decides where they await trial—at home or in jail. By law, that decision is supposed to hinge on a prediction of the defendant's risk of skipping court or reoffending. But judges

*Ludwig: University of Chicago and NBER (email: jludwig@uchicago.edu); Mullainathan: University of Chicago and NBER (email: Sendhil.Mullainathan@chicagobooth.edu); Rambachan: Massachussetts Institute of Technology (email: asheshr@mit.edu). Thanks to Nathan Hendren, Alejandro Roemer, and Josh Schwartzstein for assistance; to Paul Goldsmith-Pinkham, Greg Stoddard, Crystal Yang, and participants in our AEA session for 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.

[†]Go to https://doi.org/10.1257/pandp.20241072 to visit the article page for additional materials and author disclosure statement(s).

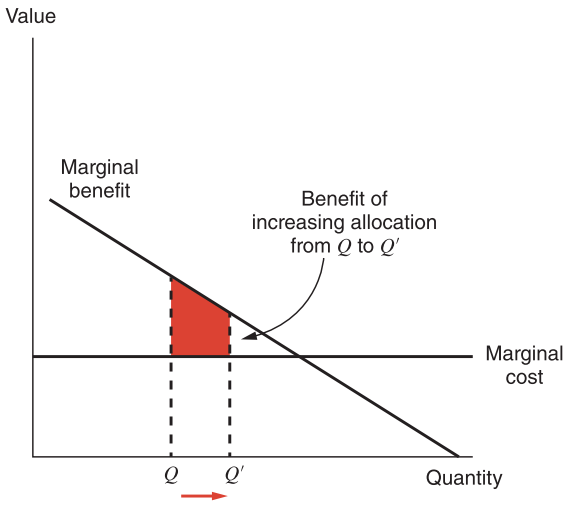

Figure 1. Stylized Illustration of the Social Welfare Gains in Increasing Allocation at Current Rankings

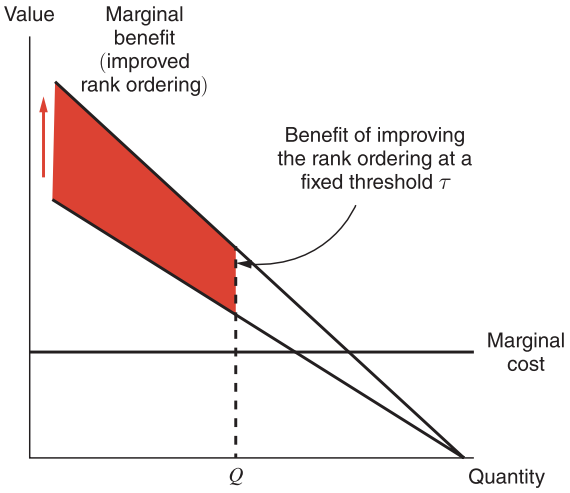

Figure 2. Stylized Illustration of the Social Welfare Gains from Algorithmic Reranking of Who Is Prioritized for Services

substantially misrank defendants, detaining many low-risk defendants and releasing many high-risk ones. Kleinberg et al. (2018) note that these predictions could instead be made by an algorithm since this application has several key ingredients: a large number of cases, a great deal of information available about each case, and a socially important decision that hinges on a prediction. Kleinberg et al. (2018) suggest that the algorithm could reduce detention rates by up to 40 percent without increasing pretrial failure rates.

Our goal here is to quantify the value of these potential gains using the social welfare metric $MVPF = \Delta W/(\Delta E - \Delta C)$, where ΔW is the value of policy impacts on affected people (willingness to pay), ΔE is the up-front change in government expenditures (e.g., to build and deploy an algorithm), and ΔC is any savings to government spending from the policy. This calculation involves at least three sources of uncertainty that are hard to quantify absent data from a deployed algorithm. The first is the algorithm's benefit. Estimating this for a hypothetical algorithm using retrospective data runs into the problem that pretrial failure can only be observed for defendants who judges release: the selective labels problem (Kleinberg et al. 2018; Rambachan 2023). Even when this can be overcome, retrospective data can't tell us anything about human compliance with a new algorithmic decision aid. A second source of uncertainty comes from the cost of building the algorithm, ΔE. This also cannot be directly quantified absent a real-world algorithm. A third source of uncertainty comes from the fact that ultimately it is the government's choice of which point to choose in the trade-off space (e.g., how much of the algorithm's gain to take from reduced detention versus reduced pretrial failures).

Despite these sources of uncertainty, progress in calculating MVPFs for algorithms is often still possible because even conservative estimates typically yield quite favorable figures. If the reduction in actual judge decisions is even only one-quarter the size of the algorithm's potential benefit (10 percent drop in detention), with 20,000 arraignments per year and something like a 50 percent release rate, the result would be 1,000 fewer arraignments. Since this could be achieved with no increase in crime, the public's willingness to pay for the algorithm—call it ΔW_P—should be nonnegative for all subgroups even if we cannot directly quantify these values. For those defendants who would have been jailed but, because of the algorithm, are released (defendants in New York are disproportionately Black and Hispanic—89 percent according to Kleinberg et al. 2018), we estimate that the value of freedom and higher labor market earnings together equal $3,200 per jail spell averted. We estimate that from 1,000 fewer detentions, the government saves $\Delta C = \$34.5$ million per year, so $MVPF = \infty$

so long as the algorithm's cost, ΔE, is below \$34.5 million.

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

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. 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 in New York City to help update the city's algorithmic decision aid to judges. We estimate that the cost of the algorithm ΔE is not more than \$4 million and find the other parameter assumptions in our policy simulation plausible as well.

II. Additional Examples

The pretrial release tool for New York City is an encouraging example and, as we will show here, not an isolated one. For example, the Office of Safety and Health Administration (OSHA) currently regulates workplace safety by targeting inspections based on the number of workplace injuries at each establishment over the past few years. Johnson, Levine, and Toffel (2023) show that an algorithm can better predict which worksites 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 and, based on the cost per serious injury and the number of days of work missed, implies a benefit to workers equal to at least $\Delta W = \$844$ million. Since the average federal tax rate for Americans is 24.8 percent, an algorithm that prevents \$844 million in lost income leads to an increase in tax revenue collection equal to $\Delta C = \$209.3$ 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.

Or consider an example from medicine: 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 them to a follow-up stress test to determine whether it is a heart attack. As Mullainathan and Obermeyer (2022) show, an algorithm that predicts patient health makes those test referrals much more accurately.

Their estimates imply that, relative to current doctor decisions, we could reduce testing by 34.7 percent with no loss in social welfare. There are 50,838 stress tests per month (cost = \$4,000) and 34,318 catheterizations per month among Medicare patients (cost = \$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 $\Delta C = \$4.8$ 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 is negative, and the MVPF of the algorithm is infinity.

As a final example, Bergman, Kopko, and Rodriguez (2023) evaluate an algorithm that screens college students into remedial (precollege) courses versus college-level courses. Underplacing students who are 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. Overplacing students can lead to them wasting time and money on classes they fail.

An RCT shows that an algorithm can predict student performance more accurately and increases placements into college-level classes by 2.6 percentage points in math and 13.6 percentage points in English (and narrows disparities across race and ethnic groups) without any decline in course pass rates. 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 saves students, on average, \$150. For the colleges in the study, the government subsidizes credit taking, and the net change in credit taking produces \$230 in savings per student, while the cost of implementing algorithmic placement is \$140 per student, so $MVPF = \$150/\$140 - \$230 = \infty$.

III. Open Questions

We have intentionally been provocative in highlighting a number of remarkably effective

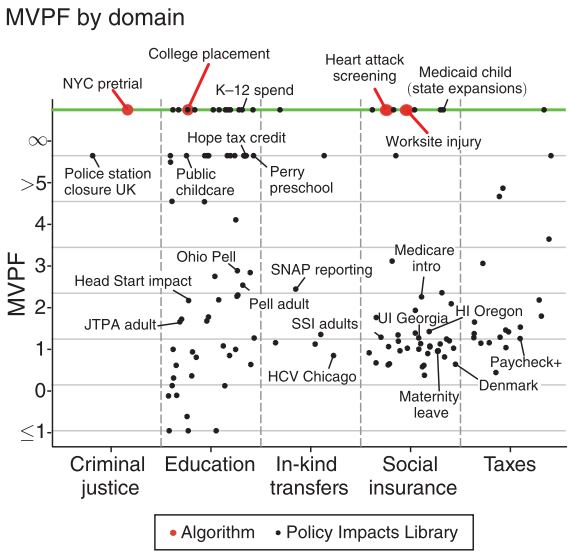

FIGURE 3. COMPARING MVPF VALUES FOR TRADITIONAL POLICIES TO THOSE FOR ALGORITHMS

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 in Figure 3 (about 15 percent of the 130 policies included in the Policy Impacts MVPF library as of this writing). Without a more exhaustive effort to calculate MVPF values for a more comprehensive set of algorithms, 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 entering more algorithms into the R&D pipeline, but we also need answers to some new economic and econometric questions that these algorithmic tools raise.

For example, 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. The result is that, for a given data frame, the algorithm is able to extract sources of signal that humans often cannot notice (Ludwig and Mullainathan 2024; Mullainathan and Rambachan 2023). Human judgment is typically such a low bar that it is easy for algorithms to soar over it.

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. 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 about which we desperately need to know more.

Relatedly, because so many algorithms are, in practice, used as decision aids, their social benefits depend on how humans make use of them. An algorithm 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. Given the diversity of findings from how humans respond to these new tools in practice, it is clear that specific design features of these algorithms might lead to variation in algorithmic impacts across settings. Much more needs to be known about how to help humans recognize their and the algorithm's sources of comparative advantage to optimally decide when to override versus follow the algorithm's predictions (e.g., Agarwal et al. 2023; Angelova, Dobbie, and Yang 2023).

IV. 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 foolproof, 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 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

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining

Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." NBER Working Paper 31422.

Angelova, Victoria, Will S. Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion." NBER Working Paper 31747.

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.

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

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

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.

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.

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

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*. https://doi.org/10.1093/qje/qjad055.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2024. "The Unreasonable Effectiveness of Algorithms." NBER Working Paper 32125.

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. 2023. "From Predictive Algorithms to Automatic Generation of Anomalies." Unpublished.

Rambachan, Ashesh. 2023. "Identifying Prediction Mistakes in Observational Data." Unpublished.