Against Predictive Optimization:
On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy

ANGELINA WANG[*], Princeton University

SAYASH KAPOOR[*], Princeton University

SOLON BAROCAS, Microsoft Research

ARVIND NARAYANAN, Princeton University

Draft: February 16, 2023. For the latest version, see
https://predictive-optimization.cs.princeton.edu/

We formalize predictive optimization, a category of **decision-making algorithms** that **use machine learning (ML)** to **predict future outcomes** of interest about **individuals**. For example, pre-trial risk prediction algorithms such as COMPAS use ML to predict whether an individual will re-offend in the future. Our thesis is that predictive optimization raises a distinctive and serious set of normative concerns that render it presumptively illegitimate. To test this, we review 387 reports, articles, and web pages from academia, industry, non-profits, governments, and modeling contests, and find many real-world examples of predictive optimization. We select eight particularly consequential examples as case studies. Simultaneously, we develop a set of normative and technical critiques that challenge the claims made by the developers of these applications—in particular, claims of increased accuracy, efficiency, and fairness. Our key finding is that these critiques apply to each of the applications, are not easily evaded by redesigning the systems, and thus challenge the legitimacy of their deployment. We argue that the burden of evidence for justifying why the deployment of predictive optimization is not harmful should rest with the developers of the tools. Based on our analysis, we provide a rubric of critical questions that can be used to deliberate or contest the legitimacy of specific predictive optimization applications.

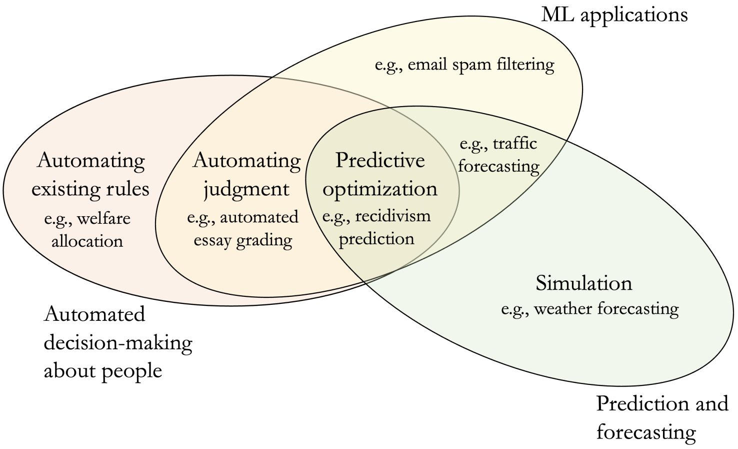
ML applications
e.g., email spam filtering
Automating existing rules
e.g., welfare allocation
Automating judgment
e.g., automated essay grading
Predictive optimization
e.g., recidivism prediction
e.g., traffic forecasting
Simulation
e.g., weather forecasting
Automated decision-making about people
Prediction and forecasting

Fig. 1. Our categorization of algorithmic decision-making systems. We focus on *predictive optimization*, the intersection of the three criteria.

[*] Equal contribution.

Electronic copy available at: https://ssrn.com/abstract=4238015

Contents

Electronic copy available at: https://ssrn.com/abstract=4238015

1 INTRODUCTION

What can go wrong with automated decision making? It depends on what exactly is being automated. In some cases, such as welfare benefits allocation, algorithms automate the process of applying a pre-existing decision-making rule. The software is designed to replace a bureaucrat, where both are expected to apply the same policy. In other cases, say resume screening, the algorithm automates the process of developing a decision-making rule in the first place. It typically uses machine learning to create a model that sorts or classifies new candidates based on patterns in historical examples. While automation is involved in both cases, the normative issues at stake are quite different. Unfortunately, distinct types of automation have tended to be conflated to the detriment of policy debates.

We focus on a subset of the second type discussed above, specifically where (1) machine learning is used to (2) make predictions about some future outcome (3) pertaining to individuals, and those predictions are used to make decisions about them. We coin the term "predictive optimization" to refer to this form of algorithmic decision-making,[1] because the decision-making rules at issue have been explicitly optimized with the narrow goal of maximizing the accuracy with which they predict some future outcome. This constrasts with manual approaches to developing decision-making rules that may involve a more deliberative process incorporating a range of considerations and goals.

Predictive optimization can be enticing. It promises to relieve policymakers of many of the difficulties in developing decision-making rules that help to realize their goals; whatever achieves the greatest possible accuracy in predicting a concrete outcome of interest is what the policy should be. It also seems to offer a more compelling and explicit justification for decision-making rules: they do not merely reflect the considered judgment of policymakers, which others might call into question as subjective; they instead seem to reflect objective patterns in the real world, which stand on their own. Presented in this manner, predictive optimization seems to remove politics from policymaking.

But in practice, predictive optimization often falls short of this ideal. Developers face a series of obstacles starting with how to make a problem amenable to predictive optimization: what should be predicted, how should this prediction inform a decision, and how does this decision help to advance the goals of the decision maker? How should one assemble the necessary data to make accurate and reliable predictions, ensure equally accurate predictions across the population, and achieve acceptable levels of accuracy given fundamental limits to prediction? And of course there are concerns around transparency and accountability: machine-learned models can be far more complex than those crafted by hand, defying meaningful inspection, and can be presented as an objective basis upon which to make decisions, concealing the many ways in which human judgment entered the process of developing them.

The thesis of this paper is that **predictive optimization fails on its own terms**. Drawing on a review of past controversies, we assemble a set of seven objections, mirroring the questions raised in the previous paragraph. By connecting these to the anatomy of predictive optimization, we argue that they are inherent and cannot be evaded without losing the essence of what makes the approach appealing in the first place. Empirically, we analyze eight case studies and show that there is either concrete or partial evidence that each of these objections applies to each of the case studies. Our critiques are either specific to predictive optimization or manifest distinctly compared to other approaches to automated decision-making (i.e, automating pre-existing rules, and machine learning of past decisions made using human judgment). For this reason, the category of predictive optimization warrants separate and careful treatment.

[1] The term is also introduced in the textbook on fairness and machine learning [14]. It was written simultaneously with this paper and shares two authors.

Electronic copy available at: https://ssrn.com/abstract=4238015

Individually, each of our critiques can threaten the perceived appropriateness of relying on predictive optimization; as a bundle, they severely undermine the legitimacy of any decision-making process based on predictive optimization. Our findings thus also serve as a diagnostic for past failures—why systems based on predictive optimization were deployed in various real-world domains but consequently discarded.

Finally, we enumerate specific questions that can help determine whether it is reasonable even to attempt to adopt predictive optimization for any given application. So far, civil society has had the burden to show that these systems are harmful. We instead suggest that for predictive optimization, the developers and decision-makers deploying these systems should have the burden of justifying why their tools are *not* harmful. Developers who cannot furnish satisfying answers to these questions should not be permitted to move forward with their proposals.

Illustrative example: driving license authorization based on risk prediction. Teenagers in many U.S. states can drive at age 16 if they pass a driving test. Teen drivers are the most dangerous of any age group. More generally, teenagers are known to engage in risky behaviors. To decrease accidents without raising the age requirement, the fictional state of West Dakota decides to use predictive optimization. There is no longer an age limit, but those applying for a license must pass not only a driving test but also a risk evaluation. This is a statistical model trained on the relationship between drivers' attributes and accidents, and predicts the probability that the applicant will be involved in an accident in the next year if granted a license. The state expects to cut accidents among teen drivers by 15% while in fact increasing the number of licensed teens by 10%.

The system does manage, initially, to cut the number of accidents and deaths. However, it also has the following consequences:

- Attributes like residential neighborhood and wealth turn out to be strong predictors of accident risk. The system thus has a disparate impact along racial and socioeconomic lines.
- People try to game the system by temporarily or permanently moving to wealthier neighborhoods. Property values in those neighborhoods go up even more, exacerbating inequality.
- The data on which the system was trained differs from the target population in many ways, including that under-16s were not in the training data. Also, since accidents are relatively rare, the developers used speeding violations as a proxy for risk. Thus, the model doesn't work as well as anticipated, and the safety improvement is much lower than expected.
- The inability to know in advance when someone will be eligible to drive causes chaos among teens' families — parents can no longer plan a vehicle purchase ahead of time or contemplate a move to a residence that necessitates a drive to school.
- The risk evaluation acquires social status over time. Those who pass it internalize the idea that they are safe drivers and respond with risk compensation behavior, nullifying the safety gains that were initially achieved.
- In contrast, those who fail the risk evaluation are stigmatized and bullied. They have no way of understanding what led to the denial—they have been told that the model uses a broad spectrum of data about their lives to predict risk, including behavior at school. While they can appeal, they cannot do so effectively since they lack an explanation of the decision.

It seems obvious that predictive optimization for driving license authorization is a bad idea, because the kinds of unintended consequences described above are easy to imagine, and the promised gains erode quickly. We argue that most other applications of predictive optimization are no different; if they don't seem as obviously problematic, it is simply because we have grown accustomed to a world in which they exist and have come to accept their drawbacks as the price of efficiency.

Electronic copy available at: https://ssrn.com/abstract=4238015

Application	Example	Uses ML?	Decisions about individuals?	Predits future outcome?	Predictive optimization?	Reason
Predictive policing. Decides geographical areas where police should be deployed.	PredPol	✓	✗	✓	✗	Decisions are not made about individuals.
Welfare allocation. Automates hand-coded rules for deciding whether an applicant is eligible for a public service, such as Medicaid.	Indiana welfare eligibility	✗	✓	✗	✗	Decisions are not made using ML. Decisions do not predict the future.
Automated essay grading. Uses past decisions by human graders to learn decision rules for grading.	TOEFL	✓	✓	✗	✗	Decisions use past judgments instead of predicting future outcomes.
Traffic prediction. Uses information about current traffic to predict estimated time of arrival.	Google Maps	✓	✗	✓	✗	Predictions are about route timings rather than an individual's outcomes.
Pre-trial risk prediction. Uses past data about individuals to predict future arrests or failure to appear in court.	COMPAS	✓	✓	✓	✓	Satisfies all three criteria for predictive optimization.

Table 1. Positive and negative examples to illustrate the definition of Predictive Optimization. An example is considered predictive optimization only if it satisfies all three criteria: it uses ML, it takes decisions about individuals, and the target variable for the ML system is a future outcome of interest.

2 PREDICTIVE OPTIMIZATION IS A DISTINCT AND IMPORTANT TYPE OF ALGORITHMIC DECISION MAKING

2.1 What is predictive optimization?

Our definition of predictive optimization has three key characteristics: (1) uses machine learning, (2) predicts future outcomes, and (3) makes decisions about individuals based on those predictions. To build an intuition for what constitutes predictive optimization, consider four hypothetical algorithms for college admissions.

(1) *A hand-coded set of rules.* The college admits applicants with a test score above a threshold and participation in at least one extracurricular activity. In Figure 1, this algorithm falls under "automating existing rules" [14]. It is not predictive optimization because it does not use ML.

(2) *An ML model trained on the past decisions made by admissions officers*, who each incorporated a range of explicit and implicit factors into their decision. In Figure 1, this algorithm falls under "automating judgment" [14]. It is not predictive optimization because it doesn't try to *predict a future event*, and rather tries to mimic past decisions and reflect their judgments.

(3) *An ML model to rank high school by predicted college performance*, and admitting students from certain high schools based on this ranking. Once again, this is not an example of predictive optimization, because it does not take decisions about *individuals*.

(4) *An ML model to predict the GPA of each applicant at the end of their first year of college based on the data in their application*, the goal being to select applicants who have a high chance of success as measured by GPA. This, finally, is an example of predictive optimization.

Electronic copy available at: https://ssrn.com/abstract=4238015

To be clear, we do not say that algorithms which fall outside these criteria are all legitimate. Rather, we want to highlight the particularly distinctive normative and technical concerns raised by applications of predictive optimization. Methodologically, predictive optimization serves as an "ideal type" [188]. This allows us to critique it as an abstract type of automated decision-making system and use these critiques to analyze concrete examples of predictive optimization. There are many ways in which the boundaries we have defined may blur in practice. Some ML models may be so simple as to resemble hand-coded rules. The predicted score may be given to a human-decision maker, who uses it as one of several factors to make the decision. The more closely an application resembles our definition, the more strongly our analysis applies.

Our hypothesis is that the category of predictive optimization is coherent enough that the same set of critiques can be levied against any application meeting the definition. At the same time, our criteria remain general enough such that they include a large number of real-world applications, as we show in Section 2.3. Table 1 provides examples of applications and clarifies where each falls under our definition.

2.2 What makes predictive optimization so compelling?

Predictive optimization has generated so much excitement and has been deployed so widely because it has many seemingly attractive characteristics—both in terms of cost and justice [163]. To understand why it is so compelling, we must ask what it is used in place of. Predictive optimization promises to improve on each of the two main traditional ways of making decisions: bureaucratic rules and human judgment.

Bureaucratic rules. Johnson and Zhang [94] argue for the superiority of predictive optimization over bureaucratic rules (they call these algorithmic prioritization and categorical prioritization respectively). Their paper is limited to government programs, but the private sector also has examples of categorical prioritization, such as organ allocation [189].[2] They describe the typical process of developing bureaucratic rules using the example of a government welfare program:

- There is a vague and generally agreed upon sense on the goals of the process, e.g., to offer assistance to those who are "deserving."
- The decision-makers select attributes that they intuitively think are relevant to the decision: income, age, number of dependent children, and criminal history.
- The attributes are discretized into categories: for instance, households are categorized as "in poverty" or not based on a threshold income.
- The attributes are combined using boolean logic to create the policy, e.g., a household qualifies if and only if it is in poverty and has some minimum number of members.

Johnson and Zhang point out many drawbacks of this approach, which we group into two main clusters. First, goals are rarely made precise, and policies often fail to meet the putative goals because their efficacy is never tested. This can be seen by how interest groups may succeed in getting their favored category added to the policy (e.g., financial assistance for veterans). Second, the resulting criteria tend to be crude because relevant attributes may not be considered, continuous variables may be thresholded, and boolean logic is not very expressive, especially compared to machine learning.

[2] More common is categorical prioritization to set eligibility requirements, such as minimum job qualifications, for a decision that is then made by human judgment.

Electronic copy available at: https://ssrn.com/abstract=4238015

This is an insightful analysis. Still, it is worth considering how serious or inevitable these drawbacks actually are. For example, for the crudeness of categorical prioritization, it turns out that across a range of tasks, well-designed numerical formulas can achieve an essentially equivalent accuracy to machine learning models while remaining straightforward to execute by hand [96]. And the lack of testing of policies is not an inherent limitation. Admittedly, there is a reason that policy evaluation is rarely carried out: it often involves causal inference and tends to be slow and expensive, yielding incomplete knowledge.

This fact motivates another notable paper that advocates for predictive optimization. Kleinberg et al. [100] argue that there is a large class of policy problems that don't require causal inference and are hence suitable for solving with machine learning using observational data. They call these "prediction policy problems"; their definition is similar to predictive optimization. Their illustrative example is as follows: you want to decide whether to carry an umbrella on a given day. Since the decision to use an umbrella doesn't impact the probability of rain, there is no causal inference required to make this decision. The only thing you need is a prediction about whether it will rain.

Their core argument is that decision problems that don't require causal inference are common and important. But there are few problems that have clear and obvious interventions once a prediction is made. That is, few real-world problems fall into the umbrella category. In our analysis, we have encountered no problems where causal modeling is unnecessary to find the best intervention, and many problems where it is clearly necessary (which includes problems that Kleinberg et al. briefly list as examples of prediction policy problems). We call this the "intervention vs. prediction" issue. In Appendix A, we present a detailed analysis of their main example: deciding whether to perform knee or hip replacement surgery.

Human judgment. The second main approach to traditional decision-making is human judgment: think of a judge making a pre-trial detention decision based on experience and intuition. Of course, judges are constrained by many bureaucratic rules, which illustrates that the boundary between the two categories is blurry. But we will continue to treat them as separate for pedagogical clarity.

Predictive optimization promises dramatic advantages over human judgment as well. Human judgment is noisy and hence inaccurate [47], costly, biased, and arbitrary in the sense that it doesn't require even the fuzzy articulation of goals and values that the bureaucratic rule-making process requires. In contrast, predictive optimization promises the consistency and efficiency of automation, and objectivity in the form of a clear target variable. This objectivity in turn promises transparency of goals and accuracy of predictions. Predictive optimization is also claimed to reduce or eliminate bias — both because of the seeming objectivity of the target variable, and because biased algorithms are claimed to be easier to fix than biased humans [125].

What about automating judgment by learning a model from past judgments? It inherits all of the limitations of human judgment except for inconsistency. Training on biased human judgments will result in a biased model. The need for training data from humans means that it cuts down on but does not eliminate the cost of human decision making. Automated judgment can be more accurate than human judgment, but is still limited in accuracy due to its training data. For example, there may be patterns in medical images that are predictive of disease and are detectable by automated systems but not by medical experts. Since these samples will be labeled negative for disease by human experts, automating judgment won't pick up on them. However, predictive optimization uses the correct ground truth based on future disease progression.

An idealized version of predictive optimization does indeed have most of these attractive qualities. But we'll question whether these conditions that are ideal for the algorithm ever arise in real use

Electronic copy available at: https://ssrn.com/abstract=4238015

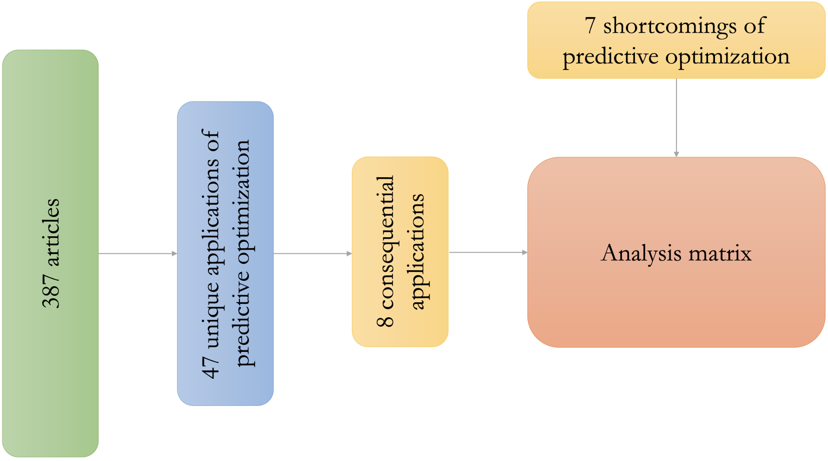
7 shortcomings of predictive optimization
387 articles
47 unique applications of predictive optimization
8 consequential applications
Analysis matrix

Fig. 2. An overview of our methods. From a corpus of 387 articles, we select eight consequential applications of predictive optimization for our case study (Section 2.4). We simultaneously select seven critiques of predictive optimization that arise as a developer is designing, creating, and deploying predictive optimization (Figure 4). We present the completed matrix in Table 2.

cases, and identify limitations that arise in realistic applications. We will also discuss other critiques that tend to be left out in the usual arguments for predictive optimization we have described.

2.3 Predictive optimization is widely deployed

So far, we have discussed the differences between predictive optimization and other decision-making processes which make it a conceptually distinct category. But why is predictive optimization important to study?

There are two reasons. First, predictive optimization is already deployed widely in consequential applications. Second, it suffers from distinct normative and technical shortcomings. We spend the rest of this section discussing our collection of consequential examples of predictive optimization and address the drawbacks of predictive optimization in the next (Section 3).

To find real-world examples of predictive optimization, we conducted a systematic search of literature from the following sources:

- **News reports:** We read articles in the top 100 search results for *algorithm* in the New York Times; sorted by relevance. To minimize U.S.-centricity, we also read articles from the first search engine results page for *algorithm* on Rest of World (an international news website that reports global tech stories).
- **Academic papers:** We read papers published in NeurIPS 2021, a prominent ML conference. We filtered relevant papers based on their title and abstract to include the ones that were related to real-world applications.
- **Kaggle competitions and datasets:** We read descriptions of Kaggle[3] competitions with a reward of $> \$50,000$, as well as the top 70 Kaggle datasets sorted by popularity.

[3] Kaggle (`www.kaggle.com`) is a data science website that hosts competitions and datasets for predictive modeling.

Electronic copy available at: https://ssrn.com/abstract=4238015

Application	Developer	What is being predicted (construct)	Proxy for the prediction (target)	Decision made based on prediction
COMPAS	Northpointe/ Equivant	Pretrial risk	Re-arrest in two years or failure to appear in court	Whether to release a defendant pre-trial or what bail amount to set
AFST	Allegheny County	Child maltreatment	Placement into foster care or multiple referrals within two years	Whether to investigate a family for child maltreatment
Hirevue	Hirevue	Job performance	Performance based on an industry-dependent performance indicator (varies by firm)	Whether to hire someone or invite them to the next round of interviews
Navigate	EAB	School dropout	Varies by school. E.g., "enrollment until next fall", "graduation within four years", or "graduation at any point of time"	Whether to offer targeted interventions to aid students
Upstart	Upstart	Creditworthiness	Repayment or future salary	Whether to offer a loan to someone and at what rates
Facebook suicide prediction	Facebook (Meta)	Suicide risk	Whether someone was assessed to be at high risk of suicide	Whether to refer someone for a welfare check
ImpactPro	Optum	Medical risk	Healthcare costs	Whether to put a patient in the high-risk health program
Velogica	SCOR	Life insurance risk	Mortality or policy lapse	Whether to offer a life insurance policy and at what rates

Fig. 3. Eight consequential applications of predictive optimization.

We also looked at an industry report [148] and two reports from non-profits [72, 95] to make sure that the applications mentioned in these reports were present in our database. This process left us with 47 potential examples of predictive optimization. Additional details about our methods and the complete list of applications are included in Appendix C.

2.4 Eight case studies of predictive optimization

We narrowed down the list of 47 examples to eight particularly consequential deployments of predictive optimization for our case studies in Section 3. We evaluated examples based on their severity (how important the decision is in an individual's life) and proliferation (how widespread the decision-making algorithm is). We limited our case studies to applications that are still in use and have sufficient documentation available for us to make informed critiques. The eight applications are outlined in Figure 3.

We compiled claims made by the developers of these systems and found three common claims: high *accuracy* in predicting the outcome, *fairness* across demographic groups, and *efficiency* gains by reducing the time spent by human decision-makers (thereby reducing costs).

Electronic copy available at: https://ssrn.com/abstract=4238015

Modeling step	Activity	Limitation	Description	Difference with automating judgment
Algorithm design	Recast decision problem as prediction problem	**Prediction vs. intervention**	Optimal predictions may not result in optimal interventions	Not formulated as prediction problem
	Operationalize construct of interest by selecting an observable proxy as the target (e.g., GPA as proxy for scholastic success)	**Target-construct mismatch**	No proxy can perfectly encapsulate construct	No target variable needed
Data collection	Select training samples collected under previous policy (e.g., students admitted in previous years; no rejected students)	**Selection bias**	Training sample doesn't match target population	Training sample includes both accepted and rejected instances
Training	Build a model to predict target variable	**Limits to prediction**	The future isn't determined yet; achievable predictive accuracy is inherently limited	Does not rely on prediction
		Disparate performance	Model may perform worse for one group or have lower rate of positive classification	Bias is an issue, but the sources and interventions tend to be different
Deployment	Make decisions using the model	**Contestability**	May be difficult due to lack of explanation of decision	Fallback to human judgment
		Goodhart's law	Decision subjects may adapt in a way that defeats goals of system	Human decision makers have some ability to notice and respond to adversarial adaptation

Fig. 4. Seven limitations of predictive optimization, mapped to the specific modeling activities that give rise to them, and brief explanations of how the automating judgment approach is different.

For instance, Hirevue's front page displays "Fast. Fair. Flexible. Finally, hiring technology that works how you want it to." Upstart claims that "future versions of the model will continue to be fair," "Upstart's model is significantly more accurate than traditional lending models," and that 73% of their loans are fully automated. These claims are used as selling points for attracting customers—for example, Optum has a document called a "sell sheet" where they list attributes such as "cost, risk and quality." We discuss developers' claims in more detail in Appendix B.

3 RECURRING SHORTCOMINGS OF PREDICTIVE OPTIMIZATION

In this section, we compile a set of seven critiques of predictive optimization. Our aim is to outline a set of objections inherent to predictive optimization that cannot be easily fixed using a design or engineering change. We generate our critiques by walking through the modeling steps involved in developing and deploying a predictive optimization application, as shown in Figure 4. Since all applications of predictive optimization involve these modeling steps, it suggests that the shortcomings we observe may apply to all of them.

3.1 Intervention vs. prediction: good predictions may not lead to good decisions

The bedrock assumption of predictive optimization is that optimal predictions lead to optimal decisions—or at least good decisions. However, since the algorithms are created using observational data, they do not directly optimize the impact of the resulting interventions [41, 64].

Electronic copy available at: https://ssrn.com/abstract=4238015

One gap between predictions and interventions is due to what is called treatment effect heterogeneity. Some individuals might be more likely to respond to an intervention compared to others, but this is not modeled by the algorithm. For example, one student may be predicted to be highly likely to drop out of school, but this may be because they are planning to move to a different city. An intervention aimed at preventing this student from dropping out would be much less effective compared to another student who might be at risk of dropping out due to underperformance.

Another gap is that decisions based on predictions might themselves affect the outcomes being predicted. For example, a higher bail amount—based on predicted recidivism—can increase the likelihood of recidivism [79]. In credit, a loan premium decided using a predictive model can negatively affect the probability of repayment. Similarly, offering different types of repayment options can change the default rate [4]. However, these effects are not modeled within the prediction problem in standard supervised ML [144].

While there are methods being developed to use machine learning to causally model and optimize decision making [10], they would require additional assumptions and/or additional data collection. Such data can often be context-specific; the plug-and-play promise of predictive optimization would no longer apply. Further, to conduct experiments to measure the effects of interventions, developers would also need to wait for years for relevant outcomes to materialize. For instance, evaluating loan repayment predictions would require information about whether the applicant defaulted after years of holding the debt.

Yet another gap is that the aggregate of individually optimal predictions may not lead to a globally optimal intervention. For instance, in a sales job, a company could prioritize hiring people based on predictions of how many sales they would close. But that doesn't account for being polite to customers or getting along with coworkers [20], which can have a long-term detrimental impact on the overall sales of the company even if the individual employee performs well. This is colloquially known as the "No Asshole Rule" [174].

Some types of interventions are more amenable to a predictive formulation than others. This restricts the scope of interventions and leads decision-makers to exclude some that could lead to more just outcomes. For example, in criminal justice, incapacitation is more amenable to a predictive formulation compared to rehabilitation [13, 176]. Alternative interventions which could decrease chances of failure to appear in court [67] are not explored in favor of a predictive approach to incapacitation. In the extreme case, the commercial pressures that favor predictive models warp the public understanding of the goals of the system [13, 57].

3.2 Target-construct mismatch: it's hard to measure what we truly care about

A *construct* is the intended outcome that the developer wants to predict. However, it is often not directly observable. Instead, the developer chooses another variable in the data that stands in as a proxy for the construct of interest. This is called the *target* variable.[4] In pre-trial risk assessment, the construct is the risk of crime or failure to appear in court if released. But crime is not directly observable, so proxies such as arrests or crime reports are used [12, 58]. Therefore, accurate prediction of re-arrest does not necessarily mean accurate prediction of re-offense. Notably, this target variable is likely to be systematically biased against Black people due to over-policing and other biases [102].

[4] The target variable is known by many names: the dependent variable, the outcome variable, the response variable, or the output variable.

Electronic copy available at: https://ssrn.com/abstract=4238015

In hiring, the construct is job performance and the target variable is an industry-dependent performance metric. For example, it could be the number of sales for a sales representative, one-year retention for flight attendants, and average client rating for tutors [108]. Job performance is famously difficult to measure [27, 154, 158], with single performance metrics receiving criticism due to the neglect of aspects such as employee behavior, e.g., politeness to customers [128] or helping their coworkers [20]. Ratings are frequently subjective [83], and specifically in the case of tutor quality ratings, student ratings of their instructors are notoriously biased [60]. All of these make efforts to predict job performance based on past data suspect.

In addition to the construct not being measurable in practice, mismatches can arise due to many other reasons. The goals of the developer may not align with that of society. For instance, the police may want to maximize the number of speeding tickets while society may want safer roads. The goals of the developer may not be precisely articulated. For example, abstract concepts like performance do not have a concrete definition. Finally, the goals of the developer may be more multi-faceted than a single target variable can capture. For example, holistic college admissions consider many different criteria.

We can rarely acquire a direct observation of the construct of interest, so to an extent, there will always be a mismatch between the target of the prediction (the measurable proxy) and the construct that it supposedly measures [92]. Target-construct mismatch contradicts developers' claim of accuracy, because no matter how good a model may be at predicting a target variable, if the target variable deviates from the construct of interest, then any claims about the accuracy fall short due to error in measurement. Additionally, if the mismatch is systematically correlated with a demographic attribute, such as race, then developers' claims of fairness are also violated.

Unobservable constructs are familiar to social scientists, who frequently use proxies to estimate their construct of interest. However, the critical difference is that the prediction setting involves working with individuals rather than aggregates, and thus the mismatch between construct and target can be far more consequential.

3.3 Distribution shifts: the training data rarely matches the deployment setting

ML methods are notorious for degradation of predictive performance under even slight changes in the distribution [71, 103]. When the distribution of data on which an ML model is trained is not representative of the distribution on which it will be deployed, model performance suffers. Thus, claims made about the model's performance might not apply to the real-world settings where it is deployed.

Distribution shifts can arise in many ways. The most common is that the training population is different from the target population in important ways, e.g., due to differences in geography. For example, the Ohio Risk Assessment System (ORAS), developed and validated using a small sample of defendants in Ohio, is used nationwide [38, 106]. In contrast, The Public Safety Assessment (PSA) tool uses a population of 1.5 million cases from 300 U.S. jurisdictions. This would seem to solve the problem, but that is still not the case. In some of the jurisdictions in which it is used, the base rate of violent recidivism is lower than the base rate in the tool's training data by more than a factor of 10. This results in risk thresholds for pre-trial detention that are severely miscalibrated, resulting in over-detention [39].

There may be situations when we only have access to data from a subset of the population, and collecting data on the entire population is hard or impossible. In this case, data from a non-representative subset of the population is used to train a model that will then be deployed on the

Electronic copy available at: https://ssrn.com/abstract=4238015

entire population of interest. This scenario is common when observational data is used: often the distribution shift arises becuase only data on successful candidates is available. Upstart's model only has access to loan default data about those that were given a loan in the first place [11]; HireVue trains their models based on a custom assessment of people already hired by the company, which is a non-random set of the population [85, 123]; Alleghany Family Screening Tool (AFST) is trained using data about public assistance, such as therapy and child welfare assistance [61], meaning that it only has access to information about those accessing public assistance, and excludes data from more well-off people who have access to private insurance and do not have to rely on public welfare.

A related issue is that developers may only be able to access data under an existing intervention. In these cases—such as data that only exists under the present criminal justice system [12]—we cannot fully evaluate the impact of introducing a decision-making algorithm in the absence of previous interventions [41]. For example, an algorithm used for predicting hospital readmission learned that patients with Asthma were at lower risk of readmission [28]. This was because patients with Asthma were more likely to be placed in an Intensive Care Unit (ICU), where they received better care which reduced their risk of readmission. Similar pitfalls are likely to arise in a vast majority of real-world applications. Chances are if a problem is important enough to deploy an algorithm to improve outcomes in a real-world setting, then we also already have an existing intervention in place for the problem (such as, sending patients with Asthma to the ICU).

Another type of selection bias is when the prediction influences the target outcome being measured and therefore creates a feedback loop [144]. For example, consider a hiring algorithm that does a poor job at assessing applicants that belong to a specific group because the training data includes few examples of people from that group. If members of this group decide to stop applying to employers who use this algorithm—perhaps because they know they are not likely to be assessed accurately—then the algorithm has even fewer examples from this population. In this case, the decision-making algorithm affects the underlying distribution being modeled. In other words, decisions cause drift.[5]

In the presence of selection bias, the performance of a decision-making algorithm cannot be measured accurately. In each of the above cases, the lack of representative data means that we cannot evaluate the performance of models trained using this data—any estimate of model performance is based on data that systematically differ from the real-world setting of interest.

3.4 Limits to prediction: social outcomes aren't accurately predictable, with or without machine learning

Predictive systems can only meet their goals, such as minimizing crime or hiring good employees, to the extent that their predictions are accurate. But there are many reasons why prediction is imperfect: both practical ones, such as limits to the ability to observe decision subjects' lives, and more fundamental ones, such as the fact that crime is sometimes a spur-of-the-moment act that can't be accurately predicted in advance.

There is accumulating evidence of strong limits to the prediction of individual-level outcomes in social systems: that is, events that are the result of social processes, compared to relatively deterministic physical or biological systems [53, 162, 192]. For example, Dressel and Farid [53]

[5] This is related to our intervention vs. prediction critique in Section 3.1. However, while here we focus on the effect of a decision on the distribution being modeled (and therefore on future decisions), earlier, we considered the effect of the decision on the outcomes of the individual about whom the decision was being made.

Electronic copy available at: https://ssrn.com/abstract=4238015

demonstrate that COMPAS is no more accurate or fair than predictions made by human participants with little or no criminal justice expertise. Further, they find that a simple linear model that only received two features is nearly equivalent to COMPAS, which had access to 137 features. Leutner et al. [108] share HireVue's predictive AUC across a range of industries, and report values from .68-.81, noting with no evidence the normative claim that "AUC values above .60 suggest the model is able to distinguish between two classes fairly well." Generally, our knowledge of the predictive performance of applications is patchy, due to a severe lack of transparency from the companies [160].

Distribution shift, discussed in the previous subsection, is also an important reason for limited accuracy. Upstart acknowledges that "There is no assurance that our AI models can continue to accurately predict loan performance under adverse economic conditions" [180]. In general, these models have been shown to perform poorly on predicting out-of-time samples [99].

Poor predictive performance undermines a common claim by developers of predictive optimization—high accuracy. These claims support the decision-making institution's ability to achieve its stated goals, for instance, protecting children at risk of maltreatment. Clearly, if the predictions were random, people would lose trust in the institution. But it is not clear what precise level of accuracy is acceptable, and this would differ between domains and applications based on many factors. What does seem to be clear is that the actual accuracy of many of these models is widely misperceived, thus misinforming this debate. For most applications, there is no consensus on what constitutes acceptable performance, and how the relative costs of false positives and false negatives should be weighed [170]. Given that people tend to trust algorithmic predictions over human judgment [111], accuracy numbers may place unfounded trust in inaccurate systems. Many decision-making systems were shut down once their actual deployed performance was revealed publicly by researchers or journalists [118, 157, 191].

It is not just the quantitative limits to prediction that matter, we also need to understand the reasons for poor accuracy, as different reasons have different moral consequences. Hardt and Kim [81] show that models in many domains are hardly more accurate than baseline models based purely on the individual's circumstances (as opposed to factors that can be attributed to the individual's agency). Lum et al. [112] show that models that predict failure to appear have large individual-level uncertainties in their risk predictions. The true risk and predicted risk often differ substantially. In other words, it's not just that these models aren't very good at achieving their instrumental aims; they also violate fairness to individuals by putting defendants with the same true risk in very different risk groups.

Limits to prediction have led to past failures of predictive optimization. Epic, one of the largest healthcare tech companies in the U.S., released a plug-and-play sepsis prediction tool in 2017. When the tool was released, the company claimed that it had an AUC between 0.76 and 0.83. Over the next five years, the tool was deployed across hundreds of U.S. hospitals. But a 2021 study found that the tool performed much worse: it had an AUC of 0.63 [191]. Following this study and a series of news reports [157], Epic stopped selling its one-size-fits-all sepsis prediction tool.

As an example of a company that went a different way, consider FICO. After finding that more complex ML models performed only slightly better than simpler regression models, it decided in favor of the latter because of interpretability [66]. Still, traditional credit scores still suffer from a number of issues [17, 36]—showing that interpretability on its own is not a panacea.

Electronic copy available at: https://ssrn.com/abstract=4238015

3.5 Disparate performance between groups can't be fixed by algorithmic interventions

Disparate performance refers to differences in performance for different demographic groups. Notably, we are not referring to differences that result from data imbalances that could be conceivably fixed through more and better data, but rather, core problems such as the fairness impossibility theorems which state that when two groups have different base rates, any calibrated algorithm cannot ensure equal false positive rates for both groups [31, 101]. For example, Angwin et al. [7] highlight the disparity in false positive rates in COMPAS, which are unavoidable due to the prioritization of the alternative fairness metrics of predictive parity, accuracy equity, and calibration [51, 53]. The numerous statistical fairness criteria each capture different moral dimensions of fairness—albeit distantly—so the impossibility theorem cannot be dealt with by declaring one metric to be the correct one.

We interpret these impossibility theorems as formalizing the well-known fact that a decision-making system that only considers people's current degree of similarities and differences, without accounting for the reasons behind those differences or histories of prejudice, will, in turn, be unjust. It may perpetuate or amplify those existing inequalities, and the cost of errors may fall disproportionately on marginalized groups. Due to the limits to prediction discussed above, the frequency of errors tends to be high.

Developers of most predictive optimization tools make claims about fairness, but rarely articulate how they define fairness or justify that choice. Fairness is a political question with many stakeholders, and there is no unbiased algorithm; only trade-offs. It is even rarer for developers or decision-makers to employ a deliberative process with input from all stakeholders including, most significantly, decision subjects [153]. After all, such deliberative processes undercut the efficiency gains that developers of predictive optimization tools tend to promise. Further, calling an automated decision-making system *fair* based on satisfying a statistical notion of fairness is misleading, since the public might equate claims of fairness with a more expansive definition. A system that is fair in a statistical sense may nonetheless perpetuate, reify, or even amplify long-standing cycles of inequality.

The relationship between predictive optimization and disparate performance is complex. On the one hand, the formalization required by predictive optimization makes discrimination more apparent [3]. The COMPAS investigation afforded a new lens through which to understand the structural racism of the criminal justice system.

On the other hand, the rigidity imposed by predictive optimization makes effective fairness interventions harder. Human decision-making is messy, but one upside is that it combines decision-making with deliberation about bias and values. This deliberation is important since fairness requires continual consensus building. The framework of algorithmic fairness presumes that consensus has been reached and scales up a single conceptualization of fairness, thus choking off one avenue for deliberation.

Predictive optimization puts the focus of interventions on the technical subsystem. Some have argued that biased algorithms are easy to fix [125], and there is a large literature on algorithmic interventions. But these interventions that mitigate statistical disparities do nothing to change the underlying conditions that gave rise to the disparities in the first place. Our view is that statistical fairness criteria are only diagnostics: the symptoms of injustice, not the disease.

The best intervention in a given situation might lie outside of the algorithm: outreach to high schools in the case of college admissions; companies investing in strengthening HBCUs in the case of workplace racial diversity; combatting sexual harassment in the workplace to improve gender

Electronic copy available at: https://ssrn.com/abstract=4238015

diversity; helping defendants show up to court rather than punishing them based on a predicted failure to appear; reconsidering money bail altogether. Such deeper interventions conflict with one of the selling points of these tools, which is technological debiasing.

Disparate performance has led to predictive optimization tools being recalled or abandoned. For example, the U.S. state of Oregon was building a tool similar to AFST. The state recalled the tool [146] after critiques of AFST were published [88], in particular, due to racial bias in the algorithm's decisions.

3.6 Providing adequate contestability undercuts putative efficiency benefits

Incorrect decisions are inevitable in practice in decision-making systems. Errors can arise in every step of an ML system—pre-processing, modeling, evaluation, deployment [109, 159]—and can lead to incorrect decisions. When decision-making algorithms are deployed in consequential settings, they must include mechanisms for contesting such decisions.

ML systems are particularly brittle and fragile to small errors, so our bar for sufficient contestability for applications of predictive optimization is accordingly high. For a decision-making algorithm to be *contestable*, there must be an explanation accompanying the decision that allows subjects to understand why they received a particular outcome. In particular, decision subjects should have access to the data about them and details about the model that was used to make the decision. In addition, there must be an accessible mechanism for reviewing and correcting contested decisions [114, 145]. This notion of contestability encompasses the related notions of providing explanations to the decision subjects and algorithmic recourse [98].

Due to the complexity of ML systems, mechanisms to offer contestability in any meaningful way would require significant overheads—in providing information about the data and models used for decision-making, educating decision subjects about the decision-making mechanism, and reviewing appeals. This contradicts developers' claims of automation and removing humans from the loop. Perhaps for this reason, none of the eight applications that we analyzed seem to provide a satisfactory level of contestability. For example, Rudin et al. [160] find that COMPAS's models have severe transparency issues and cannot be well understood even by experts. They also highlight that incorrect criminal history data has led to incorrect outputs in the past, but decision subjects cannot change or challenge the information about them that is used in the COMPAS algorithm.

A prominent example of a real-world failure of an automated system due to the lack of contestability is the Dutch welfare fraud scandal in which 30,000 parents were wrongly accused of fraud, eventually leading to the resignation of the Prime Minister and his entire cabinet [86].

3.7 Goodhart's law: predictive optimization doesn't account for strategic behavior

Goodhart's law states that "when a measure becomes a target, it ceases to be a good measure" [74]. Here, we focus on how an agent may game a metric in a way that reduces the validity of the measurement. This is also known as Goodhart's *adversarial* law [116]. A canonical example is the *cobra effect*: when the colonial British government offered bounties for dead cobras to reduce the cobra population, the response instead was people breeding more cobras to kill.

Strategic behavior is pervasive: this is what teachers do when they "teach to the test." COMPAS's inputs include defendants' agreement to sentences like "Some people must be treated roughly or beaten up just to send them a clear message" [7, 37]. Defendants can easily adapt their responses to such questions [121], which the developer acknowledges [131]. When people know that AI is used for hiring, they often use fancy words [84] or stuff their resume with the keywords from the job

Electronic copy available at: https://ssrn.com/abstract=4238015

Prediction	Case study	Intervention vs. prediction	Target-construct mismatch	Distribution shifts	Limits to prediction	Disparate performance	Lack of contestability	Goodhart's Law
Pre-trial risk	COMPAS [131]	●	●	●	●	●	●	◑
Child maltreatment	AFST [50]	●	●	●	◑	●	●	●
Job performance	HireVue [87]	◑	◑	●	◑	◑	●	●
School dropout	EAB Navigate [56, 63]	◑	◑	◑	◑	◑	●	●
Creditworthiness	Upstart [182]	◑	◑	◑	●	◑	◑	◑
Suicide	Facebook [44]	●	◑	◑	◑	◑	●	◑
Medical risk	Optum ImpactPro [136]	◑	●	●	◑	◑	◑	◑
Life insurance risk	Velogica [73]	◑	◑	◑	◑	◑	◑	◑

Table 2. A matrix with rows representing the eight consequential applications and columns representing the seven shortcomings of predictive optimization. ● represents concrete evidence that an application suffers from a limitation; ◑ represents partial or circumstantial evidence. Detailed explanations of each cell are included in Appendix B. Our key finding is that this matrix is dense.

description [164]. Attempting to improve one's credit score in a way that doesn't translate to an improved ability to repay loans is extremely common. This includes getting retail credit cards [137] and reducing the number of credit inquiries, e.g., by completing a pre-qualification form first [190]. It is natural for people to adapt their behavior to try and achieve a particular outcome, and is not just a result of people trying to "game" a system.

Strategic behavior undermines developers' claims of accuracy. Similar to target-construct mismatch, the target of prediction is no longer representative of the underlying construct of interest. Further, the distortion that leads to this drop in accuracy is not arbitrary, but systematic: those who have the access and knowledge to understand how to manipulate the system are at an advantage. Finally, some forms of strategic behavior such as obtaining and maintaining unnecessary credit cards impose a wasted (time) cost on decision subjects.

An example of a past failure due to concerns around Goodhart's Law is the LYFT score (Life Years from Transplant) for kidney allocation [153]. This score was proposed for allocating kidneys to patients in need of a transplant based on a prediction about how long they would live after the transplant [153]. Using this score would result in a disincentive for patients with kidney issues to take care of their kidney function: if their kidneys failed at a *younger* age, they would be more likely to get a transplant. That was one reason the LYFT proposal was rejected.

3.8 Summary: predictive optimization fails to meet its stated goals

We analyzed each of the 8 applications identified in Section 2.4 with respect to each of the limitations identified in Section 3. To check if an application suffers from a limitation, we synthesize past literature and, in some cases, augment this with our own analysis. We present our synthesis and analysis for all 56 cases (8 applications × 7 critiques) in detail in Appendix B, and summarize our findings in Table 2.

In the table, a filled circle (●) denotes that prior work has shown how the specific application suffers from the specific limitation. A half-filled circle (◑) denotes that either prior work has analyzed the

Electronic copy available at: https://ssrn.com/abstract=4238015

specific limitation in a closely related application or that we provide our own analysis of why the application suffers from the limitation.

Note that some rows have a high prevalence of ◑. This is due to two reasons. First, some developers are less transparent than others. When there is little to no information about an application, it is harder to understand its limitations. Second, there is a dearth of published research about some applications, perhaps as a result of the lack of transparency by the company, or perhaps simply because they are less well known to researchers. For instance, COMPAS and AFST have been extensively discussed in academic literature. As a result, we found several concrete analyses these applications. In other words, ◑ doesn't mean that evidence is weak. Rather, it highlights the need for more transparency from developers and further study of these applications.

4 CHALLENGING THE LEGITIMACY OF PREDICTIVE OPTIMIZATION

4.1 Minimum conditions for legitimacy

There have been long debates about defining legitimacy, especially in political philosophy [150, 187]. The legitimacy of a political system depends on various factors: how well it achieves its goals, whether the subjects of the political system are involved in developing the rules, and whether the decision subject has the ability to challenge decisions. Turning to automated decision-making systems, Lazar [107] argues that many ML systems are being used to exercise power over us and to govern, that the exercise of this power must meet standards of legitimacy, and that legitimacy of automated decision making requires a form of transparency ("publicity") that in turn requires explanations.

We don't adopt any specific conception of legitimacy. We only claim that, at a minimum, automated decision-making systems must fulfill their stated goals of accuracy, fairness, and efficiency as a necessary condition for legitimacy. These goals and values are central to how these systems are generally understood, and it is on the basis of these claims that the public has acquiesced to their use. If these conditions are not met, those exercising power are not being truthful to the public and are exceeding the bounds of the power they have been entrusted with.

4.2 Predictive optimization should be considered presumptively illegitimate

We started by hypothesizing that since the limitations of predictive optimization that we identify arise directly from the modeling process, they will recur across applications. The density of the matrix in Table 2 provides strong support for this hypothesis. In Section 3 we explained why none of these limitations can be evaded by technical fixes. Thus, we should expect new applications of predictive optimization to suffer from these shortcomings as well, and to fall short of the goals of accuracy, fairness, and efficiency. Therefore, the presumption should be that they fail to meet the bar for legitimacy.

Of course, it is possible that some applications are sufficiently different from the ones we have analyzed that the limitations can be overcome. Crucially, however, presumptive illegitimacy means that the onus of providing a substantive justification with respect to each of these objections must reside with the decision maker. Civil society must not have the burden of showing again and again that these systems fail to meet their goals.

This would be a major shift from the status quo. Yet, what we advocate for is not remotely radical. After all, power is *always* presumptively objectionable and requires justification [107]. Instead, what seems to have happened is that in all the hoopla around big data and AI, the shaky intellectual

Electronic copy available at: https://ssrn.com/abstract=4238015

foundations of predictive optimization were not sufficiently widely recognized, and developers' claims, especially accuracy, went unchallenged for too long, even by critics. This let predictive optimization proliferate so quickly that it appears to have become a part of the new social order, and hence normalized, to the point where challenging the entire category seems almost unthinkable. Given this state of affairs, our aim is to point out that the emperor has no clothes.

Notably, excluded from our critiques here are structural objections to the institutional context of a particular decision making system, such as calls for an overhaul of the criminal justice system [76, 119, 166]. Since structural critiques are beyond the scope of any individual decision maker, they are often not sufficiently compelling to those in charge of deployment. In contrast, our critiques are limited to those that can be seen as firmly within the scope of individual decision-makers, and thus make it hard for them to evade responsibility.

This paper focuses narrowly on how predictive optimization fails to meet its stated goals. In addition, predictive optimization systems lack other desirable properties that may be necessary for legitimacy, such as explicit deliberation and balancing of competing values. They shift authority to technocrats and away from public representatives [58, 142]. Power is often usurped by profit-driven developers at the expense of a properly authorized institution. For instance, COMPAS is built by Northpointe, a for-profit company, but is used in courts all over the United States. Government agencies often fail to recognize that when they procure automated decision making systems, they are effectively delegating policy making to the developers [127].

4.3 The need for legitimation is application-dependent

While we have argued that our critiques are broadly applicable, they are far from sufficient for a complete analysis of the legitimacy of any specific application; they are only a starting point. Application-specific considerations that we haven't considered can be critical: for instance, suicide prediction on social media threatens privacy and civil liberties, which impacts legitimacy.

Applications that are subject to our critiques aren't necessarily illegitimate. Consider a travel agency that uses ML to predict whether someone may decide to vacation soon and advertises travel deals to selected individuals. While this example falls under the scope of predictive optimization, it is less morally reprehensible than some of the applications in our list since the decision's impact is small.

Many characteristics of applications impact the legitimacy of predictive optimization.

Public or private sector: Governments derive their power—and ability to function—from being seen as legitimate. Thus, legitimacy is a critical consideration for public sector algorithms. In contrast, private firms are allowed more latitude, legally and morally, and the bar for legitimacy is lower. But there are limits. When companies make highly consequential decisions about people, they start to assume some of the power that governments do. Another way in which the distinction is blurred is when public entities such as courts use tools that have been created by private companies.

Degree of choice: In an application such as hiring, individuals have a choice because there are (typically) many companies with open positions which use different criteria for selection. On the other hand, whether someone will be placed under pre-trial detention is decided by a single entity. When the individual has less choice, the decision maker has more power over them and must therefore face more scrutiny [42].

The degree of choice is closely related to whether an application is in the public or private sector, but there are exceptions. For example, standardized tests such as the SAT and the GRE are administered by private companies but those companies are monopolies—there is usually no alternative to these tests for university applications.

Electronic copy available at: https://ssrn.com/abstract=4238015

Severity of consequences: The earlier example of travel advertisements seems relatively unobjectionable because the stakes are low. The more severe the consequences, the higher the bar for legitimacy. In general, public-sector applications such as child maltreatment prediction tend to have more serious consequences than private-sector ones, although there are many counterexamples. The degree of choice also correlates with severity.

Opportunities vs. burdens: Another factor is whether the application allocates an opportunity (e.g., admission to graduate school) or a burden (e.g., pre-trial detention). The latter raises a greater demand for legitimation. The distinction between opportunities and burdens is sometimes blurry but nonetheless conceptually useful.

4.4 A rubric to assess the legitimacy of predictive optimization

Our analysis has shown that predictive optimization applications uniformly suffer from similar shortcomings. Our key finding—that the matrix in Table 2 is dense—suggests an intervention. When applications of predictive optimization are being deployed, they can be challenged based on the critiques we compile. We provide a rubric for those trying to resist predictive optimization as well as those trying to deploy it. We detail 27 questions that must be addressed with before predictive optimization is deployed. In particular, for each of the seven shortcomings, we include 2-5 questions based on common failure modes from our analysis.

We follow and build on many efforts to critique and resist the deployment of automated decision-making systems—including those from academia [25, 33, 59, 168, 185], non-profits [72, 124], and community organizations [35, 138]. Similar concerns have been raised in past work, including unfair outcomes [61, 76], increased surveillance [35, 93, 138], false assumptions that AI systems work as intended [40, 97, 129, 130, 149, 151, 167], and questions of illegitimacy [26, 107, 184].

Through the rubric, we aim to aid civil liberties advocates, community organizers, and activists in challenging predictive optimization when deployed in their communities. When a new application is deployed, they can use the rubric to challenge developers' claims of accuracy, efficiency, and fairness that are often used to justify and legitimize the application. The rubric can also aid researchers and journalists when investigating predictive optimization by providing a concrete set of failure modes to look for in a tool. Finally, and most notably, we want to shift the burden of justifying why the tools are not harmful onto developers and decision-makers. The rubric serves as a set of basic standards that developers of predictive optimization must address in order to advertise their application as legitimate.

We present the rubric on our project website: `https://predictive-optimization.cs.princeton.edu/`.

4.5 Alternatives to predictive optimization

Suppose we have a consensus that a particular decision-making system that uses predictive optimization is illegitimate. What do we replace it with? Here, we present 4 alternatives. None are easy to implement, but we hope that together they offer a compelling vision of viable alternatives.

Note that the availability of alternatives impacts legitimacy [23]. That is, if there are no available alternatives (including shutting down the decision making system) that address the shortcomings of a predictive optimization system without causing other serious harms, the system may potentially be considered legitimate.

Electronic copy available at: https://ssrn.com/abstract=4238015

Changes to address individual critiques: To begin with, developers can address our critiques when creating decision-making algorithms. For example, causal inference techniques from program evaluation can help evaluate decisions as interventions. To ensure contestability, complex ML models could be replaced with simple, interpretable ones [96, 183], and used with institutional oversight [30, 77]. Note that each of these possible changes slightly shifts the system away from our definition of predictive optimization.

Institutional changes to address issues at their root: The need and appeal for algorithmic intervention often arises in institutions that are already suffering from crises of legitimacy, trust, and resources. For instance, predictive optimization has been adopted in hiring algorithms because existing hiring systems are broken—companies have no good way to filter through all applications for a job, and there can be hundreds of applicants for each opening. Decision-making algorithms are a band-aid to this broken system since they allow companies to reject candidates with an air of impartiality. Removing predictive optimization from this context without institutional changes does nothing to improve the root cause of failures.

Effective changes can involve redesigning institutions so that the resource in question is no longer scarce and everyone eligible can benefit. It can also include changing the institution to be less situated to adopt automated decision-making in the first place; for example, changes in the criminal justice system to rehabilitation rather than incapacitation [14].

Of course, such large-scale changes require a lot of momentum and political will, will need to be designed in a participatory fashion [169], and may take a long period of time.

Replacing decision-making systems with partial lotteries: If reducing resource scarcity is not an option, partial lotteries are worth considering [62, 78, 90, 177]. Such a lottery (for college admissions, research grants, and many other scarce resources) would select randomly among applicants who pass a minimum qualifying standard. This approach faces up to the fact that in many scenarios it is not feasible to accurately predict success or quality in advance. In contrast, predictive optimization is often used as a crutch to avoid confronting the inherent arbitrariness of the system. However, partial lotteries may not be suitable for applications where stricter definitions of equal treatment are important, such as setting bail.

Incorporating categorical prioritization: As discussed in Section 2.2, predictive optimization often replaces categorical prioritization. So one antidote to the ills of predictive optimization is to simply return to categorical prioritization or to design a hybrid approach.

Of course, categorical prioritization also suffers from several shortcomings, such as potentially prioritizing the opinions of groups who are more politically organized, as discussed by Johnson and Zhang [94]. One way to hybridize these two approaches would be to use quantitative and perhaps predictive methods to define risk categories.

Case study: COVID-19 vaccine eligibility. One application that illustrates the benefit of categorical prioritization is the rollout of COVID-19 vaccines in the U.S. in the early days when supply was scarce. A predictive optimization approach that targeted saving lives would rank individuals by the risk of death from COVID-19. But the approach that was used was categorical prioritization. To some degree, predictive considerations were used in constructing categories, notably by prioritizing groups in decreasing order of age, since age strongly predicted risk.

Categorical prioritization offered many benefits, which were collectively essential for legitimacy. First, categorical prioritization allows us to incorporate deontic considerations. Consider that healthcare workers were among the first to receive vaccines. This was not only because they were at the most risk, but as just deserts for their bravery and dedication in putting themselves in harm's

Electronic copy available at: https://ssrn.com/abstract=4238015

way for many months. Additional prioritized groups included essential workers who were keeping the economy running, and immunocompromised individuals because they were at a higher risk if they were infected. The three cases above reflect different reasons for why the individuals were prioritized. On the contrary, predictive optimization would not allow for any deliberation other than the choice of the target variable to determine eligibility.

Second, the easily understandable nature of these prioritizations allowed the public to debate the moral values underlying certain decisions. Again, public debate about the choices made using predictive optimization is hard, if not impossible, since the decision rules are opaque. Even an interpretable model such as logistic regression may be hard to understand and disseminate to the public. On the other hand, categorical prioritization relies on simple *simulatable* rules, which the public can easily keep in mind. This is necessary for productive public debate.

Simulatability is important not just for public deliberation but also for individuals. If a predictive optimization algorithm determined vaccine eligibility, each individual would get a notification when they are eligible for vaccination—they could not plan or have any idea about the criteria for eligibility.

Finally, categorical prioritization proved flexible to react dynamically to the changing reality of vaccine uptake and supply. While an early supply of vaccines was available and made eligible to certain groups, the eligibility requirements for later stages were still being debated and defined. This flexibility and adaptability was important due to the limits to prediction—during the initial stages of vaccine rollouts, neither supply nor demand could be accurately forecast. Another dimension of flexibility was that the federal government allocated vaccines to states based on population (rejecting an earlier recommendation to allocate based on risk) and allowed the states to make their own decisions regarding prioritization.

5 CONCLUSION

Previous critiques of automated decision-making have broadly fallen into two categories. The first challenges automated decision-making in a specific domain, such as criminal risk prediction [53, 80], child welfare [1], or interventions for suicide [117]. These critiques narrowly focus on harms caused by specific decision-making algorithms or those in a specific domain. While important, this work leaves us reacting post-hoc and playing Whac-a-mole to predictive optimization as it proliferates into new domains.

The second type of critique looks broadly at automated decision-making [42, 184]. Existing work in this category is less interested in any particular application of automated decision-making; the focus is on how automated decision-making systems shift power and whether they are legitimate. While also useful, this set of critiques aims so broadly that it is insufficient, on its own, to challenge the legitimacy of any specific application.

Our work offers a unifying conceptual framework to thread these two strands of the literature. Our key observation is that predictive optimization is a coherent and useful category for pointed critique. Predictive optimization is already widespread, and although the applications belong to dozens of disparate domains, they share a recurring set of shortcomings.

Our rubric translates our conceptual framework into a set of questions to contest future deployments of predictive optimization proactively. It can also be useful for developers in substantively addressing our critiques instead of making unfounded claims of accuracy, fairness, and efficiency [115].

Electronic copy available at: https://ssrn.com/abstract=4238015

ACKNOWLEDGMENTS

We are grateful to Emily Cantrell, Amrit Daswaney, Jakob Mökander, Matthew J. Salganik, and Paul Waller for their valuable feedback on drafts of this paper. We thank the participants of the Philosophy, AI, and Society workshop for their inputs. We also thank our colleagues at the Center for Information Technology Policy who provided valuable feedback during research seminars.

APPENDIX

A PREDICTION POLICY PROBLEMS ARE RARE

We revisit the application of predictive optimization to knee and hip replacement surgery allocations in Kleinberg et al. [100]. Their claim is that predictive optimization can be used to better allocate knee and hip replacement surgeries. Here we analyze how the seven shortcomings we identified might apply to their example, and show that their illustrative example might not be as straightforward as it might seem.

To begin with, we clarify our interpretation of the type of algorithm deployment proposed by the authors. If the argument is for all surgery allocation to be replaced by an algorithm, then our critiques apply in their full force to undermine the legitimacy of such an application. But if the argument is more along the lines of identifying the riskiest 1% of surgery recipients and reallocate those surgeries, then our objection is less strong. Because the precise proposal is unclear, we proceed with the former, where the algorithm has been trained to predict mortality on a patient in the 1-12 months following a hip or knee replacement surgery, in order to determine whether such a surgery should be performed.

Prediction vs intervention: Kleinberg et al. [100] predict which patients are at risk of death 1-12 months after a surgery. But they don't model who will benefit most from the surgery; instead, they assume that all patients at a given risk of mortality would benefit equally. For instance, someone who doesn't die within 12 months of surgery could still have severe complications due to the surgery. Similarly, Kleinberg et al. assume that patients would break even after a given amount of time—and pick the same threshold across patients. But different patients could recover from surgery in different time periods.

Target-construct mismatch: Kleinberg et al.'s target variable is mortality between 1-12 months. Their construct is money saved and disutility to patients. However, different patients could value surgery differently. In addition, patients might have different risk preferences. Some might be willing to take a chance despite predictions of mortality from a black-box algorithm, while others could decide that they would like to heed to model's prediction. Kleinberg et al.'s model makes these decision for the patients; in their model, they have no way of soliciting patients' feedback about the utility of surgery or their risk preferences when they are making predictions.

Distribution shift: Mullainathan and Obermeyer [126] highlight that when medical datasets are used for making predictions, they are likely to be biased. For predicting stroke in emergency department visits, they find that heavy utilization, rather than biomedical markers of stroke, turn out to be 4 of the 6 most important variables in predicting stroke. In other words, the model predicts that patients who are more likely to visit the hospital are more likely to show up with symptoms of a stroke. Patients with a longer history of visiting the hospital would be more likely to be given better care. Similar biases are likely present in electronic health records used for predicting mortality in Kleinberg et al..

Electronic copy available at: https://ssrn.com/abstract=4238015

In addition, Kleinberg et al.'s data only comes from Medicare beneficiaries, and specifically, the 1.3% of them who had a claim for hip or knee replacement surgery in 2010. This leaves out those who chose not to have a surgery or were unable to for reasons such as financial cost. Additionally, as medicine advances and different hospitals perform these surgeries [172], mortality rates and patterns are likely to change.

Limits to prediction: Kleinberg et al. state that:

> *Replacing the riskiest 10 percent with lower-risk eligibles would avert 10,512 futile surgeries and reallocate the 158 million per year (if applied to the entire Medicare population) to people who benefit from the surgery, at the cost of postponing joint replacement for 38,533 of the riskiest beneficiaries who would not have died.*

This means that for every "futile" surgery averted, about 4 people would not get a surgery who otherwise would have — based purely on an incorrect prediction about their mortality.

Without inputs from the community of people this algorithm would affect, there is no way to decide what is an acceptable threshold of false positives. But we can look to past deployments of predictive algorithms to get a sense of patients' opinions on predictive accuracy. When the LYFT score (life years from transplant) was proposed for kidney allocations to people who were likely to gain the most number of years as a result of kidney transplants, one of the reasons why the proposal was rejected was the low accuracy numbers [153]. LYFT had an AUC of 0.68, which means that the model would predict a shorter survival time for a patient who lives longer in 32% of the cases.

Kleinberg et al. don't report AUC, but it is clear that the patients and healthcare providers who will eventually be affected would have strong opinions about the threshold of accuracy that is necessary for denying someone surgery. It is far from clear that getting 80% of the decisions wrong, even if the model is only used to identify the riskiest 10% of the total surgeries, is an acceptable threshold of performance.

Disparate impact: Patients of color often receive inferior healthcare [104]. As a result, focusing on mortality risk could further benefit and allocate surgeries to those who have received better post-surgery support in the past. If the base rates between demographic groups are different, there are fundamental tradeoffs between different notions of fairness, which is not addressed by their analysis.

Contestability: The input feature has 3305 dimensions, which is nearly impossible for an individual to understand, even if the model used is logistic regression. Other factors, such as appeals and overheads due to providing patients explanations of why they weren't selected, are not factored in to their cost-benefit analysis.

Goodhart's Law: The input features include "symptoms, injuries, acute conditions, and their evolution over time." Patients are often eager to get joint replacement surgery [18] and could misreport their responses to improve their likelihood of receiving the surgery [19, 120].

B DETAILED ANALYSIS OF THE EIGHT CONSEQUENTIAL APPLICATIONS

In this section, we discuss our arguments for filling out Table 2. Each paragraph in this section corresponds to one cell in Table 2. For each application, we go over the seven critiques to see if they apply. We also briefly look at the claims made by each of the eight developers.

Electronic copy available at: https://ssrn.com/abstract=4238015

B.1 Pre-trial risk prediction in COMPAS.

Northpointe sells COMPAS as a set of risk prediction tools to be used in criminal justice settings. Here, we focus on COMPAS's pre-trial risk prediction tool. In our analysis, we rely on the company's documentation [131], as well as past literature [7, 12, 13, 38, 51, 53, 67, 79, 102, 121].

The documentation provided by COMPAS makes claims about accuracy, validity, and fairness Northpointe [131]. For instance, Northpointe claims that "...COMPAS risk scales generally fall into the moderate to good range of discrimination ability," "In overloaded and crowded criminal justice systems, brevity, efficiency, ease of administration and clear organization of key risk/need data are critical. COMPAS was designed to optimize these practical factors," and that a study from one of the tool's creators found that "COMPAS risk scales performed equally well for African American and Anglo men at discriminating recidivists in a probation sample."

The U.S. detains over 400,000 people before trial [91]. COMPAS, like many other pre-trial risk assessment tools, offers an enticing possibility: reducing the number of people detained before their trials. However, like other applications on our list, automated decision-making is a band-aid for underlying issues with the criminal justice system. For instance, over 100 civil rights groups have urged deeper changes to criminal justice institutions, such as ending pre-trial detention altogether [34]. Structural changes are outside the scope of developers of automated decision-making systems. But they would be more effective in addressing problems with the criminal justice system at their root.

(1) **Intervention vs. prediction:** Formulating pre-trial risk as a prediction problem does not do anything to increase our understanding of the underlying phenomenon, nor does it help us discover better interventions. For example, Fishbane et al. [67] find that simply providing a nudge to defendents reminding them of their court appointments is enough to reduce failures to appear. This is in addition to other potential reasons for failures to appear in court, such as financial insecurity. For example, financial insecurity could lead to defendants being unable to take leave from work in order to show up to their court appointment. Moreover, some outcomes, such as incapacitation, are more amenable to a predictive formulation, compared to others, such as rehabilitation [13]. Finally, a higher bail amount based on a high risk score can in turn increase the likelihood of recidivism [79].

(2) **Target-construct mismatch:** The construct is the risk of crime or failure to appear posed by the defendant if released. The target variable is re-arrest or failure to appear in court [12]. While the construct focuses on whether a defendant will commit a crime, the concept of "crime" is not an observable construct—for instance, not all crimes result in arrests. Therefore, the accuracy of predicting re-arrest does not translate to accuracy in measuring re-offense; in fact, there is no way to measure re-offense; the developers can only measure re-arrests. Notably, arrests are likely to be systematically biased against Black people due to known biases that result in over-policing [102].

(3) **Distribution shift:** Corbett-Davies and Goel [38] highlight sample bias in the training sets of criminal justice datasets when they are collected in one geographic area and are expected to generalize more broadly. Further, COMPAS might suffer from temporal drift in addition to geographic sample bias. While the developers of COMPAS are not transparent about the data used to train the algorithm, their documentation mentions that at least some of the data used to calibrate their models is from the years 2004-5 [131]. Finally, Bao et al. [12] highlight that the data used for pre-trial risk prediction algorithms are collected under an existing intervention—the existing criminal justice system.

Electronic copy available at: https://ssrn.com/abstract=4238015

(4) **Limits to prediction:** Dressel and Farid [53] demonstrate that COMPAS is no more accurate or fair than predictions made by human participants with little or no criminal justice expertise. Further, they find that a simple linear model that only uses two features is nearly equivalent to COMPAS, which had access to 137 features.

(5) **Disparate performance:** Angwin et al. [7] highlight the disparity in false positive rates in COMPAS, which are unavoidable due to the prioritization of the alternative fairness metrics of predictive parity, accuracy equity, and calibration [51, 53]. This disparity means that the costs of misclassification are disproportionately borne by Black defendants and communities. Given that it may not be immediately obviously which fairness criteria to be prioritized, a more in depth study into the impacts and beliefs of different groups would be needed as justification for the fairness criteria used.

(6) **Contestability:** Rudin et al. [160] find that COMPAS's models have severe transparency issues and cannot be well understood even by experts. The public does not have access to the data and model that is used to make decisions about them. Rudin et al. [160] also highlight that incorrect criminal history data has led to incorrect decisions in the past, but decision subjects cannot change or challenge the information about them that is used in the COMPAS algorithm.

(7) **Goodhart's law:** Features used in the COMPAS algorithm include survey responses asked to defendants [37]. These questions include asking for agreement to sentences like "A hungry person has a right to steal," "I have felt very angry at someone or at something," and "Some people must be treated roughly or beaten up just to send them a clear message." Social desirability bias is likely to creep in, as defendants may reasonably try to present as less likely to commit a crime [121]. Indeed, the developer of COMPAS acknowledges that some people might fake their answers in the responses to their questionnaire [131].

B.2 Child maltreatment prediction in AFST.

The Allegheny Family Screening Tool predicts which children are at risk of maltreatment to decide which households should be investigated for child maltreatment. In a 2017 report [5], the creators of AFST claimed that the tool was at least as accurate as a mammogram:

> *Measuring the accuracy of predictive tools is not simple; however, at rollout, the accuracy of the AFST was described as comparable to a mammogram: 77 percent accuracy for predicting whether a child would be placed in care within two years after being referred and screened-in for investigation, and 73 percent accuracy for predicting whether a child would be re-referred within two years after being referred and screened-out for investigation. At six-month rebuild, we intend to add an additional flag for mandatory screen-in, which is generated by a Random Forest Model which has accuracy of 88 percent (which is substantially higher than a mammogram).*

A later study by the developers of AFST [32] found that the original estimates were overoptimistic due to data leakage [32].

On fairness, the developers claimed that "fairness of algorithms is an ongoing issue for researchers in this field and the AFST research team will continue to monitor how that research impacts the AFST."

Note that while we focus on a critique of the decision-making algorithm, we acknowledge that structural changes, such as redesigning institutions to help instead of punishing families [2, 152], are overwhelmingly more important. A better algorithm is not enough to resolve fundamental issues with child welfare systems.

Electronic copy available at: https://ssrn.com/abstract=4238015

(1) **Intervention vs. prediction:** Eubanks [61] highlights that the predictive model used in AFST can end up "[producing] the outcome it is trying to measure," since being flagged by the algorithm leads to higher scrutiny and ultimately higher chances of a child being removed from their family. Additionally, AFST cannot account for the different reasons why a child could be flagged as being at risk. For instance, a child could receive a high score due to poverty or abuse. The former can be alleviated through material help to families; but the model has no way of distinguishing between the two cases.

(2) **Target-construct mismatch:** The construct is child maltreatment and the target variables are community re-referral (when two calls are made on behalf of a child within two years) and child placement (child placed in foster care within two years). Note that in this case, the data only reflects placement in foster case rather than maltreatment since there is no way to collect data on the children who are actually mistreated. There is also a mismatch between the outcomes desired by society (better treatment of children) and the outcomes desired by the algorithm's developers (higher model accuracy).

(3) **Distribution shift:** AFST does not have data about people who *do not use* public services—the data is only from those who do use these services. Parents who rely more on public services (even those unrelated to childcare) are more likely to be flagged for maltreatment. Since AFST is not trained on data from those who can access benefits privately (e.g., private mental health), it overly penalizes the poor who have to rely on public benefits [61]. As a result, the model could potentially predict poverty instead of child maltreatment. There are also disparities in the rates at which children are reported to child welfare agencies. For instance, Black children are more likely to be reported compared to other races [139].

(4) **Limits to prediction:** Salganik et al. [162] find that despite using thousands of data points about a child collected in a detailed longitudinal study, ML models could not predict outcomes about their well-being accurately. They performed about the same as simple linear models with few variables.

(5) **Disparate performance:** Cheng et al. [30] find that AFST's recommendations were more racially disparate compared to pre-AFST recommendations. The screen-in rate disparity between Black and White children based on AFST scores was 20%. While AFST creators claimed that their tool was reducing disparate impact, Cheng et al. [30] did not find evidence of this claim. Further, the creators of the tool do not explain the explicit steps taken towards reducing disparate impact.

(6) **Contestability:** Eubanks [61] outlines that human decisions and AFST decisions often disagree. Far from providing explanations to the end user, the algorithm does not even provide explanations to the humans who use the tool. As an instance of the lack of contestability leading to uncaught errors, De-Arteaga et al. [48] found that due to a glitch in the ML model, some risk scores in the AFST algorithm were misestimated. However, there was no way for the families being investigated to consult or correct these errors, since they were not aware of these issues occuring in the first place.

(7) **Goodhart's law:** Eubanks [61] finds that families being targeted as high-risk may recede from various community networks, which contributes to social isolation and parenting stress due to parents feeling like they are being watched and stigmatized. Instead of looking for greater support from community networks, the presence of the algorithm could discourage parents from seeking support with raising children. Additionally, nuisance calls by disgruntled neighbors, family members, or other potentially adversarial associates can negatively impact ASFT scores [61].

Electronic copy available at: https://ssrn.com/abstract=4238015

B.3 Job performance prediction in HireVue.

HireVue sells tools to aid hiring. This includes game-based assessments and video personality assessments, which ask candidates to answer questions which are then evaluated used an AI-based tool. HireVue earlier used to sell face analysis tools that assess candidates based on their facial expression. In 2021, after public outcry, they stopped selling this tool [6]. Our analysis here shows that fundamental issues with their tool still remain, despite claims such as "Fast. Fair. Flexible. Finally, hiring technology that works how you want it to" on their website. In another blog post, they claim that "We go on to ensure [that our] assessments actually predict job success."

(1) **Intervention vs. prediction:** The aggregate of individually optimal predictions may not lead to a globally optimal intervention. For instance, in a sales job, a company could prioritize hiring people based on potential future sales. However, hiring people who are likely to have more individual sales might not lead to better sales for the company, because it doesn't account for other behavioral factors such as being polite to customers, helping out coworkers, and working well together [20], which can have a long-term detrimental impact on the overall sales of the company even if the individual employee performs well. This is colloquially known as the "No Asshole Rule" [174].
(2) **Target-construct mismatch:** The construct is job performance and the target variable is an industry-dependent performance metric. For example, it could be the number of sales for a sales-representative, one year retention for flight attendants, and average client rating for tutors [108]. Job performance is famously difficult to measure [27, 154, 158], with single performance metrics receiving criticism due to the neglect of aspects like employee behavior, e.g., politeness to customers [128] or helping their coworkers [20]. Ratings are frequently subjective [83], and specifically in the case of tutor quality ratings, student ratings of their instructors are notoriously biased [60]. All of these make efforts to predict job performance based on past data suspect.
(3) **Distribution shift:** Hirevue trains their models based on a custom assessment. The custom assessment requires 400 people in the target job in order to train a model [123]. Thus, the model will only be trained on people already hired by the company, a non-random set of the population [85]. This can lead to a reproduction of existing population trends in the existing hired pool of employees, e.g., prioritizing men over women [46].
(4) **Limits to prediction:** Leutner et al. [108] show that AUC on some performance outcomes is .68, and specifically writes that "Notes: AUC values above .60 suggest the model is able to distinguish between two classes fairly well," which is a subjective claim without backing, given that the model would score .50 if it took decisions based on tossing a coin.
(5) **Disparate performance:** Leutner et al. [108] note that the notion of fairness adopted here is the legal one of an adverse impact ratio above 4/5. In other words, the selection rate for any protected class should not be less than 4/5ths of the rate for the group with the highest selection rate. For gender, this is defined as the female selection rate divided by that of the male selection rate. However, this legal criterion does not capture whether this model will help to intervene in cycles of workplace hiring inequity. Besides, a 20% disparity is morally problematic even if it is legally acceptable; such disparities can compound over time.
(6) **Contestability:** Candidates who are evaluated using HireVue have no insights into the criteria used for evaluation [84]. This means that they have no way to challenge incorrect or flawed decisions.

[6] https://fortune.com/2021/01/19/hirevue-drops-facial-monitoring-amid-a-i-algorithm-audit/

Electronic copy available at: https://ssrn.com/abstract=4238015

In addition, HireVue did not share details about how well the model performed. Before 2020, there were no public audits of their hiring assessment tools [153]. They released an audit of their tool in 2020. However, before downloading the audit, you need to accept a restrictive agreement on using or sharing the report's findings:

> *By downloading this document you acknowledge and agree this report is the sole and exclusive intellectual property of HireVue, Inc., and you agree you shall not use, copy, excerpt, reproduce, distribute, display, publish, etc. the contents of this report in whole, or in part, for any purpose not expressly authorized in writing by HireVue, Inc.*

Since the findings in the report cannot be shared, let alone interrogated, we do not consider it a public audit.

(7) **Goodhart's law:** When people know that AI is used for hiring, they often do not understand how it works and use fancy words to optimize for their success [84] or stuff their resume with the keywords in the job description [164]. Whether or not this works, it affects the behavior of job candidates. Another report found that wearing glasses and adding a bookshelfl in your video intervieiw makes your automated interview scores higher [82].

B.4 School dropout prediction in EAB Navigate.

EAB Navigate is a tool for colleges to target interventions at students who are claimed to be at risk of dropping out [63]. The company uses personal information, academic performance, app activity, and credit trends from students to predict if they will drop out of school. It is meant to be used by college administrators to evaluate whom to target with interventions.

On accuracy, in a document from the company detailing the tool (found and uploaded by The Markup [63]), the company boasts:

> *The performance of your institution's Student Success Predictive Model has been extensively optimized and evaluated; the model will provide your school and its advisors with invaluable and otherwise unobtainable insight into your students' likelihood of academic success. The model incorporates the latest breakthroughs in statistics and data science and places your institution at the cutting edge of student-insight technology. Your advisors may use it with confidence to both assess individual students and design effective and efficient targeted campaigns.*

On efficiency, they state on their website: "Navigate's workflow solutions help academic advisors, faculty, and other staff scale interventions."

On fairness, when confronted with the racial disparities in their system, an EAB employee said: "What we are trying to do with our analytics is highlight these disparities and prod schools to take action to break the pattern."

(1) **Intervention vs. prediction:** Predictions are often for specific students who seem likely to drop out, but it isn't clear that the types of interventions the school has in mind, e.g., offering advising to particular students, is what is needed by each student. For example, sometimes school-wide structural interventions are more helpful [54, 55, 147], and other times, individuals from specific populations, e.g., LGBTQ students, would benefit from a more catered type of intervention than general advising [52]. Predictions don't help us understand which interventions would help individual students.

(2) **Target-construct mismatch:** The construct is attrition and the target variable is an observable metric of dropping out of school. EAB Navigate uses different target variables for each school—for example, *enrollment until next fall*, *graduation within 4 years*, and *graduation at any point of*

Electronic copy available at: https://ssrn.com/abstract=4238015

time [63] are all used to model attrition. While it can be seen as a desirable feature that the model can be customized for the needs of each school, this also tells us there are numerous reasonable options for how to operationalize school dropout. Each value-laden choice of target variable has its own implications for target-construct mismatch [110], and there is no clear justification to explain that the choice of a particular target variable is more than arbitrary. For example, if the target variables is *enrollment next fall*, then this does not count the student who dropped out for a family emergency, who took a semester off to work, or a reason unrelated to school success itself. Even in the academic literature, the definition of dropout or attrition varies greatly [21, 141], making comparisons between studies hard [140].

(3) **Distribution shift:** The training data for these models come from students who are already enrolled in the institution, as the filtering criterion frequently includes qualifications such as "had at least one registered term" and "were seeking a degree" [63]. As the distribution of students attending colleges change over time, the model might not be reliable on new students or those who do not fit the criteria of a typical college student. However, that EAB Navigate trains individual models for each institution does alleviate part of this concern to some extent.

(4) **Limits to prediction:** The developers take no inputs from the different stakeholders about acceptable thresholds of accuracy at which students should be referred for an intervention. There is no deliberation or consensus process for feedback from students. In fact, students on whom the system was used were unaware that their performance was tracked using predictive models [63].

(5) **Disparate performance:** The model targets Black students much more compared to their white counterparts. Feathers [63] find that “[a]t the University of Massachusetts Amherst, for example, Black women are 2.8 times as likely to be labeled high risk as White women, and Black men are 3.9 times as likely to be labeled high risk as White men.” The definition of fairness used is not specified by the company, nor do they justify why their model is fair.

(6) **Contestability:** Students being evaluated using EAB Navigate have little to no information about the models being used [63]—in many the cases, the students do not even know that the model is being used to make predictions about them. This forecloses any chance of contestability. Further, the data and model used for making predictions is not available to the students, so any inaccuracies are left undetected and unchallenged.

(7) **Goodhart’s law:** To lower dropout rates, one school (Mt. Saint Mary’s) has previously used dropout predictions to preemptively kick out students so their graduation rates stay high [175]. Other schools have pushed students, especially Black students, out of majors like science and math into ones with lower dropout rates [63].

B.5 Creditworthiness prediction in Upstart.

Upstart is a lending platform which predicts potential lenders’ creditworthiness. They use over 1,600 datapoints about customers, and claim that “future versions of the model will continue to be fair,” “Upstart’s model is significantly more accurate than traditional lending models,” and that 73% of their loans are fully automated.

The Consumer Financial Protection Bureau (CFPB) issued a no-action letter for Upstart after checking that their model did not suffer from disparate performance across demographic groups. This letter provided Upstart special immunity:

Electronic copy available at: https://ssrn.com/abstract=4238015

> *The CFPB had granted special regulatory treatment to Upstart by immunizing the lender from being charged with fair lending law violations with respect to its underwriting algorithm, while the "no-action letter" remained in force.*[7]

In July 2022, Upstart issued an update to their model and did not wait for regulatory approval from CFPB. As a result, the bureau rescinded their earlier no-action letter for Upstart.

(1) **Intervention vs. prediction:** Research research has shown that offering different types of repayment options can change the default rate [4]. However, algorithms for predicting whether someone will repay their loans cannot take this into account owing to their predictive formulation.

(2) **Target-construct mismatch:** The construct is creditworthiness. While Upstart does not specify the target variable, various sources indicate it could be future repayment of loans or future salary [105, 155, 182]. For future repayment, a cutoff date will necessarily have to be set to collect the data, which will not differentiate between someone who pays their loan back right after this cutoff date, and someone who never pays it back. For future salary, this may have little bearing on whether a loan will actually be repaid, because an individual could choose to spend their income in other ways.

(3) **Distribution shift:** Upstart's model, like other credit models, only has access to loan default data about those that were given a loan in the first place [11]. As a result, their data sample only has information about a portion of the population.

(4) **Limits to prediction:** Upstart [180] acknowledges that "There is no assurance that our AI models can continue to accurately predict loan performance under adverse economic conditions." In general, these models have been show to perform poorly on predicting out-of-time samples when the economic conditions have changed drastically [99].

(5) **Disparate performance:** One of Upstart's main claims is it uses alternative data sources like which college an applicant goes to as an indicator of how likely they are to pay back a loan. There are findings that the algorithm is biased against Historically Black Colleges and Universities [171]. Upstart responded by claiming that the report contains mistakes [181], and points to the CFPB report which found that Upstart had increased acceptance rates and decreased APRs across all protected groups compared to existing models [65]. However, there is no transparency on the accuracy across these different groups, including whether it is higher for one group than another.

    Note that even under the original no-action letter, CFPB only compared Upstart's model with traditional alternatives such as FICO scores. This is a weak bar for fairness: legal compliance cannot break long running cycles of financial inequality [45, 49].

    Finally, as we have discussed (Section 3.5), when the base rates differ across groups, there is a fundamental tension between different notions of fairness. Upstart does not detail how it considers fairness in its decision-making algorithm, what fairness tradeoffs it adopts, or how these decisions were made in the first place.

(6) **Contestability:** Banks and credit unions which use Upstart provide adverse action notices to loan applicants. In addition, applicants who are denied a loan due to incorrect documentation can upload proof of qualification to correct the documentation. However, Upstart does not specify the model used to make decisions, so the precise criteria used for making decisions are still unknown to applicants. This means that applicants cannot challenge the decision-making algorithm on grounds of being incorrect or unfair. Moreover, their model uses over 1,600

[7] `https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-order-to-terminate-upstart-no-action-letter/`

Electronic copy available at: https://ssrn.com/abstract=4238015

features, which makes it hard to challenge incorrect data or to understand how the model is used for making decisions.

(7) **Goodhart's law:** Upstart uses credit scores as an input feature. There are several actions you can take that will increase your credit score, but are not necessarily related to your ability to repay a loan. This includes getting retail credit cards [137] and reducing the number of credit inquiries, e.g., by completing a pre-qualification form first [190].

B.6 Suicide prediction on Facebook.

Facebook uses suicide prediction algorithms to direct help to users who are at risk. One of the company's blog posts details how they do this [44]. We use information from this post as well as several academic works on suicide prediction [29, 117] in our analysis. While Facebook claims that their random forests model outperforms previous attempts at suicide prediction, they do not provide any concrete accuracy numbers or other details about their model.

(1) **Intervention vs. prediction:** Facebook's algorithm only focuses on predicting suicides, and does not try to understand what interventions might be useful. In particular, Facebook's wellness checks prompt visits from police—this can be harmful to the decision subjects. People with untreated mental illness are sixteen times more likely to be killed in encounters with the police [179]. Committing suicide is also illegal in a lot of countries [117].
(2) **Target-construct mismatch:** The construct is risk of suicide and the target variable is user reports of their peers and subsequent actions taken by content moderators [117]. Facebook can only measure specific outcomes such as reports of suicidal content from an individual's friends, and has no way to measure the underlying construct. As we will see in our analysis of Goodhart's law below, user reports are often weaponized for mass reporting on Facebook.
(3) **Distribution shift:** The model is trained based on comments and reports from Facebook friends of users [117]. This data can be biased towards users who have a higher number of friends, geographic areas where reporting suicidal content is not taboo, and languages in which Facebook's content moderation is more active, such as English [43].
(4) **Limits to prediction:** There are no performance metrics reported about the suicide prediction model used by Facebook. Peer-reviewed research on suicide prediction is based on murky evidence and suffers from lack of construct validity [29]. As a result, we have no evidence of how well Facebook's tools work.
(5) **Disparate performance:** Without any data released about who is referred for welfare checks by Facebook, it is hard to investigate disparate impact of the decisions. However, the models used by Facebook rely on natural language processing techniques [117] which have been shown to be biased in multiple studies [24, 173].
(6) **Contestability:** Facebook users have no way to interrogate what data or model leads to wellness checks for suicide prevention [117]. They have no way to opt-out of such checks. Further, since Facebook is a private company, there is no accountability for their model validation and data collection: unlike research conducted at universities, which requires approval from the IRB, Facebook only has an optional internal ethics review, and the final decision about produce launches is based on the company's discretion.
(7) **Goodhart's law:** Content moderation policies such as Facebook's suicide prediction can lead to a change in user behaviour in multiple ways. First, users can be targetted as victims of mass reporting. For example, Buzzfeed reporter Katie Notopolous has her account banned due to mass-reporting [133]. Similar tactics could be applied to other benign accounts by reporting their content for suicidal content. Second, people have been shown to change their behavior

Electronic copy available at: https://ssrn.com/abstract=4238015

online when they know that they are being surveilled [143]. This could also affect how people discuss self-harm and suicidal thoughts on platforms like Facebook.

B.7 Medical risk prediction in Optum ImpactPro.

Optum's algorithm uses past history of a patient to predict whether they are at high medical risk. The rationale is that providing pre-emptive care to high risk patients can reduce costs in the long run, for instance, by reducing visits to the emergency room. Optum's ImpactPro software came under scrutiny after a prominent *Science* paper by Obermeyer et al. [134] found evidence of racial bias. Optum has a document called a "sell sheet"[8] where they list attributes such as "cost, risk and quality" as their main selling points. It goes on to claim that:

> *With Impact Pro, you can determine which individuals are in need of specialized intervention programs and which intervention programs are likely to have an impact on the quality of individuals' health. ... This rich insight helps you manage populations proactively, prioritizing timely interventions to enhance the clinical and financial returns on your population health management programs.*

Note that while our focus is on automated decision-making in healthcare, we acknowledge the need for structural changes. For instance, the need for Optum's algorithm only arises because of resource limitations. It is worth questioning whether these limitations are necessary in the first place. Adopting structural changes would address the root causes of algorithmic failures instead of merely addressing their symptoms with algorithmic band-aids [69].

(1) **Intervention vs. prediction:** Healthcare is a long-term goal. Preventative healthcare may be more effective than reacting to individuals who are already at a high risk [161]. Intervening when a patient is already high-risk may not be as effective as other interventions earlier on. In addition, a predictive formulation does not address the problem of higher mistrust that Black patients have for the healthcare system [6, 8]. In other words, predictions about current risk status do not inform which interventions are most effective for patients.

(2) **Target-construct mismatch:** The construct is healthcare needs and the target variable is healthcare costs [134]. However, due to reasons such as unequal access to healthcare, the costs are often a poor proxy for the actual healthcare needs. Instead, the higher rate of White patients who are labeled as high risk illustrates existing inequities in the healthcare system. Obermeyer et al. say that one of the reasons why Optum could be using healthcare costs as a proxy is because the construct of health needs is hard to operationalize.

(3) **Distribution shift:** s discussed in Appendix A, Mullainathan and Obermeyer [126] find that electronic health records are biased towards people who show up to the hospital more often. It is possible that patients who are more likely to show up to hospitals will be marked as being at higher risk. Meanwhile, patients who hesitate to visit the hospital, or those who cannot afford frequent visits, will be underrepresented in the dataset.

    In addition, the model only has access to data from Black patients under existing inequities in healthcare. Due to higher distrust of physicians by Black people, they have less hospital visits for illnesses of similar severities [6, 8], and they often spend lesser compared to White people with similar levels of health issues [134].

    Optum also doesn't disclose how it deals with shifts in geographic distributions. A model trained on a nationally representative sample is unlikely to perform well at a local level; similarly, there are differences in how people access healthcare in rural vs. urban locations [186].

[8] https://www.optum.com/content/dam/optum3/optum/en/resources/sell-sheet/impact-pro-sell-sheet.pdf

Electronic copy available at: https://ssrn.com/abstract=4238015

(4) **Limits to prediction:** The prediction of similar tasks such as triage prioritization have been shown to have strong limits to prediction, among physicians, nurses, and computer programs [22]. Optum's white paper on their ImpactPro model claims an R^2 value of 0.295 for their cost-based risk prediction model [135]. It is far from clear that this meets the bar of high accuracy for such a consequential decision about patients, and Optum had no measures in place to test whether patients who would eventually be classified using this model accept the model's performance. As we have seen with LYFT (Appendix A; Robinson [153]), patients and healthcare providers can have strong opinions on what accuracy threshold is acceptable.

(5) **Disparate performance:** Optum does not provide evidence of any notion of fairness being satisfied by their model. Obermeyer et al. [134] highlight alternative target variables that can lead to more calibrated models. But given that different demographic groups have different base rates, all decision-making systems will be subject to difference in performance [9]—regardless of the choice of target variable.

(6) **Contestability:** Patients cannot challenge the outcomes from Optum's algorithm. One mechanism to provide some level of recourse is that while patients above the 97th percentile of risk are directly enrolled in the high-risk program, those above the 55th percentile of risk are referred to their primary care provider; this can provide some amount of recourse to the patients who are not directly enrolled in the high-risk program based on the algorithm's recommendations.

(7) **Goodhart's law:** If patients are classified as high-risk, they get access to better healthcare and in some cases as assigned a 1-1 healthcare provider [70]. Patients are often eager to get these perks, since they provide better and higher priority care [178]. In a state with bad existing healthcare, patients have many incentives to misrepresent their conditions to get access to better healthcare [165].

B.8 Life insurance risk prediction in Velogica.

Velogica is a life insurance pricing tool from the company SCOR. On its webpage[9], Velogica makes the following claims (emphasis ours):

- **90% of underwriting evaluations within one minute**
- 24/7 application submission capability
- Thousands of applications processed weekly
- Sales process and decision that takes less than 15 minutes
- **Human underwriting on less than 5% of applications**
- **Validated risk assessment effectiveness**
- Increased application flow
- Underwriting decisions in less than a minute
- Lower acquisition costs
- Lower underwriting costs
- Consistent underwriting assessments
- Success across distribution channels
- Risk participation of a life reinsurance leader

(1) **Intervention vs. prediction:** Similar to creditworthiness, changing the payment structure could help decrease the probability that a given customer will lapse. An intervention focused on helping a given customer avoid lapse would be more useful than predicting who is most likely to lapse and charging them a higher premium. However, life insurance companies profit from

[9] https://www.scorgloballifeamericas.com/en-us/solutions/us/Pages/US-Velogica.aspx

Electronic copy available at: https://ssrn.com/abstract=4238015

lapsed policies and lose money on those who keep their policies; this leads to higher premiums being charged to customers upfront and increases the probability of lapse [75]. That is, higher premiums lead to a feedback loop; they put people at higher risk of lapse.

(2) **Target-construct mismatch:** The construct is policy risk posed by an applicant and the target is mortality or policy lapse due to lack of payment. In many cases, individuals might lapse on paying insurance premiums because of lack of information about when or how to pay [132] or forgetfulness [75], rather than any change in their ability to pay or their mortality risk. This would indicate higher risk but occurs only due to lack of knowledge or information about how to pay life insurance premiums and indicates a mismatch between the construct (policy risk) and its operationalization in the form of mortality or policy lapse. In some cases, life insurance companies might *expect* policies to lapse since they can benefit from the upfront costs paid by a customer [75]. In this case, modeling policy lapse as the target variable is a valid design choice for the company, but diverges from the notion of "risk-based insurance" as it is commonly understood. In addition, it shows how the company's goals (profitability) diverge from society's goals (widespread access to insurance).

(3) **Distribution shift:** The model only has access to data from people who were underwritten using a life insurance policy in the first place, and might not generalize to atypical populations or those who are currently underserved.

(4) **Limits to prediction:** Velogica has not released any public details about the accuracy of their models. While Velogica publicly claims that its model has been validated [73], there is no information about how the validation was carried out or how well the model performs.

(5) **Disparate performance:** Velogica uses several features which are correlated with sensitive attributes such as race. For example, they use criminal history as a feature in their model. This is correlated with race and could lead to worse outcomes for Black people. Black people also have higher mortality rates compared to White people [68]. This difference in base rates implies that a calibrated model cannot have equal false positive rates across races. In this context, the definition of fairness used is a normative question. Far from engaging with this question, Velogica does not provide any information about its definition of fairness or data about outcomes by race.

The history of life-insurance is marked by unequal treatment of Black people. In many cases, insurers did not allow Black people to purchase policies, or if they did, only gave them insurance for funeral money [113]. In this context, treating insurance as a problem of statistical disparity does nothing to rectify long-standing cycles of racial disparity.

(6) **Contestability:** When life insurance companies have used automated underwriting, they have found that resolving appeals is harder because there is no human who made the initial decision who can be consulted about their logic for adverse actions [15].

(7) **Goodhart's law:** One of the items Velogica uses to determine insurance rates is answers to a self-reported questionnaire, which can be easily gamed by candidates to get favorable premiums. Life insurance companies, including Velogica, use results from medical exams and lab tests to price life insurance policies [122]. Many websites advise candidates who are getting tested in a medical exam for life insurance on how to perform better, for example, by not consuming nicotine or tobacco 12-24 hours before an exam [89, 156].

C INVENTORY OF PREDICTIVE OPTIMIZATION ALGORITHMS

On our project website, we present 47 potential applications of predictive optimization collected from our list of 387 articles, Kaggle contests, and datasets. To code each source, we selected each application of decision-making algorithms mentioned in it. For example, a New York Times article

Electronic copy available at: https://ssrn.com/abstract=4238015

published in 2020 is titled “Is an Algorithm Less Racist Than a Loan Officer?”. Based on this article, we included automated loan decisions as one example of a decision-making algorithm. Then, we checked to see if this application met any of our three criteria: (1) a real-world example of such a prediction exists; (2) a Kaggle dataset or contest exists for this prediction task; (3) the prediction task is referenced by an academic paper as a potential application. If it met any of these criteria, we assessed whether the application fit into our definition of predictive optimization, and added it to our list.

The first two authors conducted the coding. Each source was initially coded by either one of the authors. Once all sources were coded, both authors went over the list and came to a consensus about the coding decisions. The first set of collection and coding was done during March 2022, with a second pass done in February 2023.

Some applications, notably recommender systems, don’t fit precisely under our definition of predictive optimization, but we decided to include them. Much of our analysis still applies to this category, although it mainly addresses individual harms, whereas it is the structural harms that are more important [16].

We release our list of 47 applications and their corresponding sources on our website `https://predictive-optimization.cs.princeton.edu/`. We note that this compilation is a work-in-progress and we welcome suggestions for new applications.

Electronic copy available at: https://ssrn.com/abstract=4238015

REFERENCES

[1] J. Khadijah Abdurahman. 2021. Calculating the Souls of Black Folk: Predictive Analytics in the New York City Administration for Children's Services. *Columbia Journal of Race and Law* 11, 4 (July 2021), 75–110. https://doi.org/10.52214/cjrl.v11i4.8741

[2] J. Khadijah Abdurahman. 2022. Birthing Predictions of Premature Death. https://logicmag.io/home/birthing-predictions-of-premature-death/

[3] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 252–260.

[4] Kwami Adanu and Emma Boateng. 2015. Predicting loan repayment default among second tier borrowers in Ghana. *International Journal of Entrepreneurship and Small Business* (2015).

[5] Allegheny county analytics. 2017. Allegheny Family Screening Tool: Frequently asked questions. https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/07/AFST-Frequently-Asked-Questions.pdf

[6] Marcella Alsan and Marianne Wanamaker. 2018. Tuskegee and the Health of Black Men. *The Quarterly Journal of Economics* 133, 1 (Feb. 2018), 407–455. https://doi.org/10.1093/qje/qjx029

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[8] Katrina Armstrong, Karima L. Ravenell, Suzanne McMurphy, and Mary Putt. 2007. Racial/ethnic differences in physician distrust in the United States. *American Journal of Public Health* 97, 7 (July 2007), 1283–1289. https://doi.org/10.2105/AJPH.2005.080762

[9] Deepshikha Charan Ashana, George L. Anesi, Vincent X. Liu, Gabriel J. Escobar, Christopher Chesley, Nwamaka D. Eneanya, Gary E. Weissman, William Dwight Miller, Michael O. Harhay, and Scott D. Halpern. 2021. Equitably Allocating Resources during Crises: Racial Differences in Mortality Prediction Models. *American Journal of Respiratory and Critical Care Medicine* 204, 2 (July 2021), 178–186. https://doi.org/10.1164/rccm.202012-4383OC

[10] Susan Athey. 2017. Beyond prediction: Using big data for policy problems. *Science* 355, 6324 (Feb. 2017), 483–485. https://doi.org/10.1126/science.aal4321

[11] J. Banasik, J. Crook, and L. Thomas. 2003. Sample Selection Bias in Credit Scoring Models. *The Journal of the Operational Research Society* 54, 8 (2003), 822–832. https://www.jstor.org/stable/4101652

[12] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (Dec. 2021). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract-round1.html

[13] Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, and Jonathan Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. *arXiv:1712.08238 [cs, stat]* (July 2018). http://arxiv.org/abs/1712.08238 arXiv: 1712.08238.

[14] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

[15] Mike Batty, David Moore, and Mike McCarty. 2010. Automated Life Underwriting: Phase 2. https://www.soa.org/globalassets/assets/Files/Research/Projects/research-auto-life-underwriting-2.pdf

[16] Claire Benn and Seth Lazar. 2022. What's Wrong with Automated Influence. *Canadian Journal of Philosophy* 52, 1 (Jan. 2022), 125–148. https://doi.org/10.1017/can.2021.23

[17] Laura Blattner and Scott Nelson. 2021. *How Costly is Noise? Data and Disparities in Consumer Credit.* Technical Report arXiv:2105.07554. arXiv. https://doi.org/10.48550/arXiv.2105.07554 arXiv:2105.07554 [cs, econ, q-fin] type: article.

[18] Ashley W. Blom, Neil Artz, Andrew D. Beswick, Amanda Burston, Paul Dieppe, Karen T. Elvers, Rachael Gooberman-Hill, Jeremy Horwood, Paul Jepson, Emma Johnson, Erik Lenguerrand, Elsa Marques, Sian Noble, Mark Pyke, Catherine Sackley, Gina Sands, Adrian Sayers, Victoria Wells, and Vikki Wylde. 2016. *Understanding patient's experiences of total hip and knee replacement: a qualitative study.* NIHR Journals Library. https://www.ncbi.nlm.nih.gov/books/NBK379631/

[19] Brandon L. Boring, Kaitlyn T. Walsh, Namrata Nanavaty, Brandon W. Ng, and Vani A. Mathur. 2021. How and Why Patient Concerns Influence Pain Reporting: A Qualitative Analysis of Personal Accounts and Perceptions of Others' Use of Numerical Pain Scales. *Frontiers in Psychology* 12 (July 2021), 663890. https://doi.org/10.3389/fpsyg.2021.663890

[20] Walter Borman and S. Motowidlo. 1993. Expanding the Criterion Domain to Include Elements of Contextual Performance. *Personnel Selection in Organizations* (Jan. 1993), 71–98. https://digitalcommons.usf.edu/psy_facpub/1111

Electronic copy available at: https://ssrn.com/abstract=4238015

[21] Alex J. Bowers, Ryan Sprott, and Sherry A. Taff. 2012. Do We Know Who Will Drop Out? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity, and Specificity. *The High School Journal* 96, 2 (2012), 77–100. https://www.jstor.org/stable/23351963

[22] Judith C Brillman, David Doezema, Dan Tandberg, David P Sklar, Kathleen D Davis, Shelby Simms, and Betty J Skipper. 1996. Triage: Limitations in Predicting Need for Emergent Care and Hospital Admission. *Annals of Emergency Medicine* 27, 4 (April 1996), 493–500. https://doi.org/10.1016/S0196-0644(96)70240-8

[23] Allen Buchanan and Robert O. Keohane. 2006. The Legitimacy of Global Governance Institutions. *Ethics & International Affairs* 20, 4 (Dec. 2006), 405–437. https://doi.org/10.1111/j.1747-7093.2006.00043.x

[24] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[25] Ryan Calo. 2021. Modeling Through. *Duke Law Journal* (Oct. 2021). https://papers.ssrn.com/abstract=3939211

[26] Ryan Calo and Danielle Keats Citron. 2021. The Automated Administrative State: A Crisis of Legitimacy. *Emory Law Journal* (2021).

[27] J. Campbell. 1990. Modeling the performance prediction problem in industrial and organizational psychology. *Handbook of industrial/organizational psychology* (1990). http://www.sciepub.com/reference/136816

[28] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Sydney NSW Australia, 1721–1730. https://doi.org/10.1145/2783258.2788613

[29] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine* 3, 1 (March 2020), 1–11. https://doi.org/10.1038/s41746-020-0233-7

[30] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3491102.3501831

[31] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *FATML* (2016). http://arxiv.org/abs/1703.00056 arXiv: 1703.00056.

[32] Alexandra Chouldechova, 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*. PMLR, 134–148. https://proceedings.mlr.press/v81/chouldechova18a.html

[33] Danielle Keats Citron. 2008. Technological Due Process. *Washington University Law Review* 85, 6 (Jan. 2008), 1249–1313. https://openscholarship.wustl.edu/law_lawreview/vol85/iss6/2

[34] CivilRights.org. 2018. Pretrial Risk Assessments. https://civilrights.org/edfund/pretrial-risk-assessments/

[35] Stop LAPD Spying Coalition. 2021. AUTOMATING BANISHMENT: The Surveillance and Policing of Looted Land. https://stoplapdspying.org/automating-banishment-the-surveillance-and-policing-of-looted-land/

[36] Consumer Financial Protection Bureau. 2012. Analysis of Differences between Consumer- and Creditor-Purchased Credit Scores. *SSRN Electronic Journal* (2012). https://doi.org/10.2139/ssrn.3790636

[37] Contributed by Julia Angwin (ProPublica). 2011. Sample COMPAS Risk Assessment: COMPAS "CORE". https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE

[38] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]* (Aug. 2018). http://arxiv.org/abs/1808.00023 arXiv: 1808.00023.

[39] Ethan Corey. 2019. How a Tool to Help Judges May Be Leading Them Astray. https://theappeal.org/how-a-tool-to-help-judges-may-be-leading-them-astray/

[40] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2022. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. https://doi.org/10.48550/arXiv.2206.14983 arXiv:2206.14983 [cs].

[41] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 582–593. https://doi.org/10.1145/3351095.3372851

[42] Kathleen Creel and Deborah Hellman. 2021. *The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems*. SSRN Scholarly Paper ID 3786377. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3786377

Electronic copy available at: https://ssrn.com/abstract=4238015

[43] Elizabeth Culliford and Brad Heath. 2021. Language Gaps in Facebook's Content Moderation System Allowed Abusive Posts on Platform: Report. https://thewire.in/tech/facebook-content-moderation-language-gap-abusive-posts

[44] Dan Muriello, Lizzy Donahue, Danny Ben-David,, Umut Ozertem, and Reshef Shilon. 2018. Under the hood: Suicide prevention tools powered by AI. https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

[45] William Darity Jr., Darrick Hamilton, Mark Paul, Alan Aja, Anne Price, Antonio Moore, and Caterina Chiopris. 2018. What We Get Wrong About Closing the Racial Wealth Gap. http://narrowthegap.org/images/documents/Wealth-Gap---FINAL-COMPLETE-REPORT.pdf

[46] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[47] Robyn M. Dawes, David Faust, and Paul E. Meehl. 1989. Clinical Versus Actuarial Judgment. *Science* 243, 4899 (March 1989), 1668–1674. https://doi.org/10.1126/science.2648573

[48] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[49] Ellora Derenoncourt, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick. 2022. Wealth of Two Nations: The U.S. Racial Wealth Gap, 1860-2020. https://doi.org/10.3386/w30101

[50] DHS. 2019. Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions. https://www.alleghenycountyanalytics.us/index.php/2019/05/01/developing-predictive-risk-models-support-child-maltreatment-hotline-screening-decisions/

[51] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe* (2016).

[52] Donna J. Dockery. 2012. School Dropout Indicators, Trends, and Interventions for School Counselors. *Journal of School Counseling* 10, 12 (2012). https://eric.ed.gov/?id=EJ978868

[53] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580 Publisher: American Association for the Advancement of Science.

[54] Mark Dynarski, Linda Clarke, Brian Cobb, Jeremy Finn, Russell Rumberger, and Jay Smink. 2008. Dropout Prevention: A Practice Guide. *Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education* (Sept. 2008).

[55] Mark Dynarski and Philip M Gleason. 2002. How Can We Help? What We Have Learned From Recent Federal Dropout Prevention Evaluations. *Journal of Education for Students Placed at Risk (JESPAR)* (2002).

[56] EAB Navigate. 2022. Navigate | Student Success Management System | EAB. https://eab.com/products/navigate/

[57] Jessica Eaglin. 2019. Technologically Distorted Conceptions of Punishment. *97 Washington University Law Review 483 (2019)* (Jan. 2019). https://www.repository.law.indiana.edu/facpub/2862

[58] Jessica M. Eaglin. 2017. Constructing Recidivism Risk. *Articles by Maurer Faculty* (2017).

[59] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment. *Criminal Justice and Behavior* 46, 2 (Feb. 2019), 185–209. https://doi.org/10.1177/0093854818811379 Publisher: SAGE Publications Inc.

[60] Justin Esarey and Natalie Valdes. 2020. Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education* (Feb. 2020). https://www.tandfonline.com/doi/abs/10.1080/02602938.2020.1724875

[61] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, New York, NY.

[62] Ferric C. Fang and Arturo Casadevall. 2016. Research Funding: the Case for a Modified Lottery. *mBio* 7, 2 (May 2016), e00422–16. https://doi.org/10.1128/mBio.00422-16

[63] Todd Feathers. 2021. Major Universities Are Using Race as a "High Impact Predictor" of Student Success. https://themarkup.org/machine-learning/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success

[64] Carlos Fernández-Loría and Foster Provost. 2021. Causal Decision Making and Causal Effect Estimation Are Not the Same... and Why It Matters. *INFORMS Journal on Data Science* (Sept. 2021). http://arxiv.org/abs/2104.04103 arXiv:2104.04103 [cs, stat].

[65] Patrice Alexander Ficklin and Paul Watkins. 2019. An update on credit access and the Bureau's first No-Action Letter. https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter/

Electronic copy available at: https://ssrn.com/abstract=4238015

[66] FICO Decisions. 2018. Machine Learning and FICO Scores. (2018). https://www.fico.com/en/resource-access/download/6559

[67] Alissa Fishbane, Aurelie Ouss, and Anuj K. Shah. 2020. Behavioral nudges reduce failure to appear for court. *Science* 370, 6517 (Nov. 2020), eabb6591. https://doi.org/10.1126/science.abb6591

[68] Anna Flagg. 2021. The Black Mortality Gap, and a Document Written in 1910. *The New York Times* (Aug. 2021). https://www.nytimes.com/2021/08/30/upshot/black-health-mortality-gap.html

[69] Alison P. Galvani, Alyssa S. Parpia, Abhishek Pandey, Charlotte Zimmer, James G. Kahn, and Meagan C. Fitzpatrick. 2020. The imperative for universal healthcare to curtail the COVID-19 outbreak in the USA. *eClinicalMedicine* 23 (June 2020). https://doi.org/10.1016/j.eclinm.2020.100380

[70] Ishani Ganguli, E. John Orav, Eric Weil, Timothy G. Ferris, and Christine Vogeli. 2018. What Do High-Risk Patients Value? Perspectives on a Care Management Program. *Journal of General Internal Medicine* 33, 1 (Jan. 2018), 26–33. https://doi.org/10.1007/s11606-017-4200-1

[71] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (Nov. 2020), 665–673. https://doi.org/10.1038/s42256-020-00257-z

[72] Michele Gilman. 2020. Poverty Lawgorithms. https://datasociety.net/library/poverty-lawgorithms/

[73] SCOR Global. 2022. SCOR Global Life. https://www.scorgloballifeamericas.com:443/en-us/solutions/us/Pages/US-Velogica.aspx

[74] C. A. E. Goodhart. 1984. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice: The UK Experience*. Macmillan Education UK, London, 91–121. https://doi.org/10.1007/978-1-349-17295-5_4

[75] Daniel Gottlieb and Kent Smetters. 2021. Lapse-Based Insurance. *American Economic Review* 111, 8 (Aug. 2021), 2377–2416. https://doi.org/10.1257/aer.20160868

[76] Ben Green. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 594–606. https://doi.org/10.1145/3351095.3372869

[77] Ben Green. 2022. The Flaws of Policies Requiring Human Oversight of Government Algorithms. https://doi.org/10.2139/ssrn.3921216

[78] Kevin Gross and Carl T. Bergstrom. 2019. Contest models highlight inherent inefficiencies of scientific funding competitions. *PLOS Biology* 17, 1 (Jan. 2019), e3000065. https://doi.org/10.1371/journal.pbio.3000065

[79] Arpit Gupta, Christopher Hansman, and Ethan Frenchman. 2016. *The Heavy Costs of High Bail: Evidence from Judge Randomization*. SSRN Scholarly Paper 2774453. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.2774453

[80] Bernard E. Harcourt. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press, Chicago, IL. https://press.uchicago.edu/ucp/books/book/chicago/A/bo4101022.html

[81] Moritz Hardt and Michael P. Kim. 2022. Backward baselines: Is your model predicting the past? https://doi.org/10.48550/arXiv.2206.11673 arXiv:2206.11673 [cs, stat].

[82] Elisa Harlan and Oliver Schnuck. 2021. Objective or Biased. https://interaktiv.br.de/ki-bewerbung/en/

[83] Melissa Hart. 2006. *Subjective Decisionmaking and Unconscious Discrimination*. SSRN Scholarly Paper ID 788066. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=788066

[84] Drew Harwell. 2019. A face-scanning algorithm increasingly decides whether you deserve the job. *Washington Post* (2019). https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

[85] James J. Heckman. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47, 1 (1979), 153–161. https://doi.org/10.2307/1912352

[86] Melissa Heikkilä. 2022. Dutch scandal serves as a warning for Europe over risks of using algorithms. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

[87] HireVue. 2022. HireVue Hiring Platform: Video Interviews, Assessment, Scheduling, AI, Chatbot. https://www.hirevue.com/

[88] Sally Ho and Garance Burke. 2022. An algorithm that screens for child neglect raises concerns. https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1

[89] Cameron Huddleston. 2022. How To Get A Better Rate On An Existing Life Insurance Policy If Your Health Has Improved. https://www.forbes.com/advisor/life-insurance/rate-reconsideration/

[90] Hundred.org. 2017. Student Government Lotteries. https://hundred.org/en/innovations/student-government-lotteries#ad460767

[91] Prison Policy Initiative. 2023. Pretrial Detention. https://www.prisonpolicy.org/research/pretrial_detention/

Electronic copy available at: https://ssrn.com/abstract=4238015

[92] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[93] Diego Jemio, Alexa Hagerty, and Florencia Aranda. 2022. The Case of the Creepy Algorithm That 'Predicted' Teen Pregnancy. *Wired* (2022). https://www.wired.com/story/argentina-algorithms-pregnancy-prediction/

[94] Rebecca Ann Johnson and Simone Zhang. 2022. What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in U.S. Social Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1671–1682. https://doi.org/10.1145/3531146.3533223

[95] Divij Joshi. 2021. AI Observatory. https://ai-observatory.in/

[96] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2020. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183, 3 (June 2020), 771–800. https://doi.org/10.1111/rssa.12576

[97] Frederike Kaltheuner. 2021. *Fake AI*. Meatspace Press. https://shop.meatspacepress.com/products/fake-ai-e-book

[98] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899

[99] Hua Kiefer and Tom Mayock. 2020. Why Do Models That Predict Failure Fail? *Federal Deposit Insurance Corporation Working Paper Series* (2020).

[100] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction Policy Problems. *American Economic Review* 105, 5 (May 2015), 491–495. https://doi.org/10.1257/aer.p20151023

[101] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)* (2017). http://arxiv.org/abs/1609.05807 arXiv: 1609.05807.

[102] Dean Knox, Will Lowe, and Jonathan Mummolo. 2020. Administrative Records Mask Racially Biased Policing. *American Political Science Review* 114, 3 (Aug. 2020), 619–637. https://doi.org/10.1017/S0003055420000039 Publisher: Cambridge University Press.

[103] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 5637–5664. https://proceedings.mlr.press/v139/koh21a.html

[104] Gary L. Kreps. 2006. Communication and Racial Inequities in Health Care. *American Behavioral Scientist* 49, 6 (Feb. 2006), 760–774. https://doi.org/10.1177/0002764205283800

[105] Katja Langenbucher and Patrick Corcoran. 2022. Responsible AI Credit Scoring – A Lesson from Upstart.com. *De Gruyter* (2022).

[106] Edward J. Latessa, Richard Lemke, Matthew Makarios, and Paula Smith. 2010. Creation and Validation of the Ohio Risk Assessment System (ORAS) | Office of Justice Programs. *Federal Probation* 74, 1 (June 2010), 16–22. https://www.ojp.gov/ncjrs/virtual-library/abstracts/creation-and-validation-ohio-risk-assessment-system-oras

[107] Seth Lazar. 2022. Legitimacy, Authority, and the Political Value of Explanations. https://doi.org/10.48550/arXiv.2208.08628 arXiv:2208.08628 [cs].

[108] Kiki Leutner, Josh Liff, Lindsey Zuloaga, and Nathan Mondragon. 2021. HireVue's Assessment Science. *HireVue White Paper* (Oct. 2021). https://webapi.hirevue.com/wp-content/uploads/2021/11/2021_10_HireVue_Assessment_Science_white_paper-FINAL-1.pdf?_ga=2.65347438.1736480487.1648481172-1018073685.1646944864&_gac=1.93515503.1648481197.CjwKCAjwuYWSBhByEiwAKd_n_ozIeWkJyt84zksCiwZuKvz7c1ZWBhxvqGRE7fcwAZTZGmbkSO9PgBoC-dwQAvD_BwE

[109] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (Dec. 2021). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html

[110] Lydia T. Liu, Serena Wang, Tolani Britton, and Rediet Abebe. 2022. Lost in Translation: Reimagining the Machine Learning Life Cycle in Education. *PNAS* (Sept. 2022). https://doi.org/10.48550/arXiv.2209.03929 arXiv:2209.03929 [cs].

[111] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (March 2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Electronic copy available at: https://ssrn.com/abstract=4238015

[112] Kristian Lum, David B. Dunson, and James Johndrow. 2021. Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment. https://doi.org/10.48550/arXiv.2102.01135 arXiv:2102.01135 [stat].

[113] Lynette Hazelton and Oscar Perry Abello. 2022. What's a Black life worth to insurance companies? https://www.inquirer.com/news/inq2/more-perfect-union-life-insurance-history-racism-20221129.html

[114] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–25. https://doi.org/10.1145/3449180

[115] Gianclaudio Malgieri and Frank A. Pasquale. 2022. From Transparency to Justification: Toward Ex Ante Accountability for AI. https://doi.org/10.2139/ssrn.4099657

[116] David Manheim and Scott Garrabrant. 2019. Categorizing Variants of Goodhart's Law. *arXiv:1803.04585 [cs, q-fin, stat]* (Feb. 2019). http://arxiv.org/abs/1803.04585 arXiv: 1803.04585.

[117] Mason Marks. 2019. Artificial Intelligence-Based Suicide Prediction. *YALE JOURNAL OF HEALTH POLICY, LAW, AND ETHICS* 18:3 (2019), 24.

[118] Paris Martineau. 2022. Toronto Tapped Artificial Intelligence to Warn Swimmers. The Experiment Failed. https://www.theinformation.com/articles/when-artificial-intelligence-isnt-smarter

[119] Sandra G. Mayson. 2017. Dangerous Defendants. https://doi.org/10.2139/ssrn.2826600

[120] Harald Merckelbach, Brechje Dandachi-FitzGerald, Daniel van Helvoort, Marko Jelicic, and Henry Otgaar. 2019. When Patients Overreport Symptoms: More Than Just Malingering. *Current Directions in Psychological Science* 28, 3 (June 2019), 321–326. https://doi.org/10.1177/0963721419837681

[121] Jeremy F. Mills and Daryl G. Kroner. 2005. An Investigation Into the Relationship Between Socially Desirable Responding and Offender Self-Report. *Psychological Services* (2005). https://psycnet.apa.org/record/2005-06059-007

[122] Cindy Mitchell and Peter Komsthoeft. 2019. Underwriting Innovation: Harnessing the Differences. https://www.scorgloballifeamericas.com:443/en-us/knowledgecenter/underwriting-innovation-harnessing-the-differences

[123] Nathan Mondragon, Josh Liff, Kiki Leutner, and Lindsey Zuloaga. 2021. Assessments Overview and Implementation. *HireVue White Paper* (Oct. 2021).

[124] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

[125] Sendhil Mullainathan. 2019. Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times* (Dec. 2019). https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

[126] Sendhil Mullainathan and Ziad Obermeyer. 2017. Does Machine Learning Automate Moral Hazard and Error? *American Economic Review* 107, 5 (May 2017), 476–480. https://doi.org/10.1257/aer.p20171084

[127] Deirdre K. Mulligan and Kenneth A. Bamberger. 2019. Procurement As Policy: Administrative Process for Machine Learning. *Berkeley Technology Law Journal* 34 (2019). https://escholarship.org/uc/item/90t9k477

[128] K. Murphy and L. Kroeker. 1988. Dimensions of Job Performance. https://doi.org/10.21236/ada194951

[129] Alexander Martin Mussgnug. 2022. The predictive reframing of machine learning applications: good predictions and bad measurements. *European Journal for Philosophy of Science* 12, 3 (Aug. 2022), 55. https://doi.org/10.1007/s13194-022-00484-8

[130] Arvind Narayanan. 2019. How to recognize AI snake oil. (2019), 21.

[131] Northpointe. 2019. A Practitioner's Guide to COMPAS Core. (April 2019). http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf

[132] Mira Norton, Liz Hamel, and Mollyann Brodie. 2014. Assessing Americans' Familiarity With Health Insurance Terms and Concepts. https://www.kff.org/health-reform/poll-finding/assessing-americans-familiarity-with-health-insurance-terms-and-concepts/

[133] Katie Notopoulos. 2017. How Trolls Locked My Twitter Account For 10 Days, And Welp. https://www.buzzfeednews.com/article/katienotopoulos/how-trolls-locked-my-twitter-account-for-10-days-and-welp

[134] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453. https://doi.org/10.1126/science.aax2342

[135] Optum. 2021. Guiding population health management programs with comprehensive clinical insight. https://www.optum.com/content/dam/optum3/optum/en/resources/white-papers/wf4785959-impact-pro-white-paper.pdf

Electronic copy available at: https://ssrn.com/abstract=4238015

[136] Optum. 2022. Health Risk Analytics - Impact Pro. https://www.optum.com/business/health-plans/data-analytics/predict-health-risk.html

[137] Bev O'Shea. 2022. How to Score Points in the Credit Game. https://www.nerdwallet.com/article/finance/how-to-score-points-in-the-credit-game

[138] Our Data Bodies. 2022. AMC 2022 Preview: A Look into the Abolish Carceral Tech Track. https://www.odbproject.org/2022/06/27/amc-2022-preview-a-look-into-the-abolish-carceral-tech-track-2/

[139] Vincent J. Palusci and Ann S. Botash. 2021. Race and Bias in Child Maltreatment Diagnosis and Reporting. *Pediatrics* 148, 1 (July 2021), e2020049625. https://doi.org/10.1542/peds.2020-049625

[140] Robert J. Panos and Alexander W. Astin. 1968. Attrition Among College Students. *American Educational Research Journal* 5, 1 (Jan. 1968), 57–72. https://doi.org/10.3102/00028312005001057

[141] Timothy J. Pantages and Carol F. Creedon. 1978. Studies of College Attrition: 1950-1975. *Review of Educational Research* 48, 1 (1978), 49–101. https://doi.org/10.2307/1169909

[142] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[143] Jon Penney. 2016. Chilling Effects: Online Surveillance and Wikipedia Use. https://papers.ssrn.com/abstract=2769645

[144] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. *ICML* (2020). http://arxiv.org/abs/2002.06673 arXiv: 2002.06673.

[145] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–52. https://doi.org/10.1145/3411764.3445315

[146] The Associated Press. 2022. Oregon is dropping an artificial intelligence tool used in child welfare system. *NPR* (June 2022). https://www.npr.org/2022/06/02/1102661376/oregon-drops-artificial-intelligence-child-abuse-cases

[147] Frances Prevatt and F. Donald Kelly. 2003. Dropping out of school: A review of intervention programs. *Journal of School Psychology* 41, 5 (Sept. 2003), 377–395. https://doi.org/10.1016/S0022-4405(03)00087-6

[148] PricewaterhouseCoopers. 2017. PwC's Global Artificial Intelligence Study: Sizing the prize. https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

[149] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

[150] John Rawls. 1993. *Political Liberalism*. Columbia University Press.

[151] Benjamin Recht. 2022. Machine Learning has a validity problem. http://benjamin-recht.github.io/2022/03/15/external-validity/

[152] Dorothy E. Roberts. 2022. *Torn apart: how the child welfare system destroys black families–and how abolition can build a safer world* (first edition ed.). Basic Books, New York.

[153] David G. Robinson. 2022. *Voices in the Code: A Story about People, Their Values, and the Algorithm They Made*. Russell Sage Foundation, New York, NY.

[154] David Robotham and Richard Jubb. 1996. Competences: measuring the unmeasurable. *Management Development Review* 9, 5 (Jan. 1996), 25–29. https://doi.org/10.1108/09622519610131545

[155] Andrew Rombach. 2018. Upstart CEO Dave Girouard Talks Machine Learning, AI, and Loans. https://lendedu.com/blog/upstart-ceo-dave-girouard-talks-machine-learning-artificial-intelligence-personal-loans/

[156] Georgia Rose. 2022. Life Insurance Medical Exams: What to Expect. https://www.nerdwallet.com/article/insurance/life-insurance-medical-exams

[157] Casey Ross. 2022. Epic's overhaul of a flawed algorithm shows why AI oversight is a life-or-death issue. https://www.statnews.com/2022/10/24/epic-overhaul-of-a-flawed-algorithm/

[158] Maria Rotundo and Paul R. Sackett. 2002. The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology* (2002). https://psycnet.apa.org/record/2002-00102-006

[159] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[160] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review* 2, 1 (Jan. 2020). https://doi.org/10.1162/99608f92.6ed64b30

Electronic copy available at: https://ssrn.com/abstract=4238015

[161] Michael Sagner, Amy McNeil, Pekka Puska, Charles Auffray, Nathan D. Price, Leroy Hood, Carl J. Lavie, Ze-Guang Han, Zhu Chen, Samir Kumar Brahmachari, Bruce S. McEwen, Marcelo B. Soares, Rudi Balling, Elissa Epel, and Ross Arena. 2017. The P4 Health Spectrum – A Predictive, Preventive, Personalized and Participatory Continuum for Promoting Healthspan. *Progress in Cardiovascular Diseases* 59, 5 (March 2017), 506–521. https://doi.org/10.1016/j.pcad.2016.08.002

[162] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117, 15 (April 2020), 8398–8403. https://doi.org/10.1073/pnas.1915006117

[163] Dario Sansone and Anna Zhu. 2021. Using Machine Learning to Create an Early Warning System for Welfare Recipients. *Institute of Labor Economics* (2021).

[164] Hilke Schellmann. 2022. Finding it hard to get a new job? Robot recruiters might be to blame. *The Guardian* (May 2022). https://www.theguardian.com/us-news/2022/may/11/artitifical-intelligence-job-applications-screen-robot-recruiters

[165] Shelly K. Schwartz. 2010. When Patients Lie to You. https://www.roswellpark.org/partners-in-practice/white-papers/when-patients-lie-you

[166] Katia Schwerzmann. 2021. Abolish! Against the Use of Risk Assessment Algorithms at Sentencing in the US Criminal Justice System. *Philosophy & Technology* 34, 4 (Dec. 2021), 1883–1904. https://doi.org/10.1007/s13347-021-00491-2

[167] Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, and E. Glen Weyl. 2021. How AI Fails Us. *Justice, Health, and Democracy Impact Initiative* (2021). https://ethics.harvard.edu/files/center-for-ethics/files/howai_fails_us_2.pdf?m=1638369605

[168] Eric Silver and Lisa L. Miller. 2002. A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control. *Crime and Delinquincy* 48, 1 (2002).

[169] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a Design Fix for Machine Learning. *Workshop at International Conference on Machine Learning* (Aug. 2020). http://arxiv.org/abs/2007.02423 arXiv: 2007.02423.

[170] Megan T. Stevenson and Sandra G. Mayson. 2022. Pretrial Detention and the Value of Liberty. *Virginia Law Review* 108, 3 (May 2022). https://www.virginialawreview.org/articles/pretrial-detention-and-the-value-of-liberty/

[171] Student Borrower Protection Center. 2020. Educational Redlining. *Student Borrower Protection Center* (2020). https://protectborrowers.org/wp-content/uploads/2020/02/Education-Redlining-Report.pdf

[172] Adarsh Subbaswamy and Suchi Saria. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics (Oxford, England)* 21, 2 (April 2020), 345–352. https://doi.org/10.1093/biostatistics/kxz041

[173] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[174] Robert Sutton. 2010. *The No Asshole Rule: Building a Civilized Workplace and Surviving One That Isn't*. Business Plus, New York.

[175] Susan Svrluga. 2016. University president allegedly says struggling freshmen are bunnies that should be drowned. *Washington Post* (2016). https://www.washingtonpost.com/news/grade-point/wp/2016/01/19/university-

Electronic copy available at: https://ssrn.com/abstract=4238015

president-allegedly-says-struggling-freshmen-are-bunnies-that-should-be-drowned-that-a-glock-should-be-put-to-their-heads/

[176] Christopher T. Lowenkamp, Marie VanNostrand, and Alexander Holsinger. 2013. The Hidden Costs of Pretrial Detention. *Arnold Foundation* (2013).

[177] The British Academy. 2022. BA/Leverhulme Small Research Grants. https://www.thebritishacademy.ac.uk/funding/ba-leverhulme-small-research-grants/

[178] Ingrid Torjesen. 2022. Covid-19: Incomplete lists of vulnerable patients left many unprotected, desperate, and afraid. *BMJ* 376 (Feb. 2022), o528. https://doi.org/10.1136/bmj.o528

[179] Treatment Advocacy Center. 2015. Overlooked in the Undercounted. https://www.treatmentadvocacycenter.org/overlooked-in-the-undercounted

[180] United States Securities and Exchange Commission. 2021. *Form 10-K for Upstart Holdings, Inc.* Technical Report. https://www.sec.gov/ix?doc=/Archives/edgar/data/1647639/000164763922000009/upst-20211231.htm

[181] Upstart Blog. 2018. Upstart's Commitment to Fair Lending. https://www.upstart.com/blog/upstarts-commitment-to-fair-lending

[182] Upstart Blog. 2020. Introducing the Credit Decision API for Banks. https://www.upstart.com/blog/introducing-credit-decision-api

[183] VantageScore. 2017. Scoring Credit Invisibles: Using machine learning techniques to score consumers with sparse credit histories. (2017). https://vantagescore.com/wp-content/uploads/2022/02/20171009_Machine-Learning-online-3.pdf

[184] Ari Waldman. 2019. Power, Process, and Automated Decision-Making. *Fordham Law Review* 88, 2 (Nov. 2019), 613. https://ir.lawnet.fordham.edu/flr/vol88/iss2/9

[185] Ari Ezra Waldman. 2020. Algorithmic Legitimacy. In *The Cambridge Handbook of the Law of Algorithms*, Woodrow Barfield (Ed.). Cambridge University Press, Cambridge, 107–120. https://doi.org/10.1017/9781108680844.005

[186] Robin Warshaw. 2017. Health Disparities Affect Millions in Rural U.S. Communities. https://www.aamc.org/news-insights/health-disparities-affect-millions-rural-us-communities

[187] Max Weber. 1919. Politics as a Vocation. *The Vocation Lectures* (1919).

[188] Max Weber. 1949. Max Weber on the methodology of the social sciences. (1949).

[189] David L. Weimer. 2007. Public and private regulation of organ transplantation: liver allocation and the final rule. *Journal of Health Politics, Policy and Law* 32, 1 (Feb. 2007), 9–49. https://doi.org/10.1215/03616878-2006-027

[190] Alexandria White. 2019. 6 easy tips to help raise your credit score. https://www.cnbc.com/select/easy-tips-to-help-raise-your-credit-score/

[191] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA internal medicine* 181, 8 (Aug. 2021), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

[192] Tal Zarsky. 2016. The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values* 41, 1 (Jan. 2016), 118–132. https://doi.org/10.1177/0162243915605575

Electronic copy available at: https://ssrn.com/abstract=4238015