Foundations of Prediction in the Public Sphere

Unai Fischer Abaigar

München 2026

Foundations of Prediction in the Public Sphere

von

Unai Fischer Abaigar

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der
Ludwig-Maximilians-Universität München
eingereicht am 08. Juni, 2026

1. Gutachter: Prof. Dr. Christoph Kern
2. Gutachter: Prof. Dr. Avi Feller
3. Gutachter: Prof. Dr. Julia Stoyanovich

Tag der Disputation: 27.07.2026

Acknowledgements

This dissertation would not exist without the support and guidance of my advisors, Frauke Kreuter and Christoph Kern. I am deeply grateful to both of them for taking a chance on a physics student looking to find his footing in a new field, and for the space they gave me to grow into a researcher. Frauke taught me to always reach for the next big challenge, to find genuine joy in building an academic community, and to see the value of terseness. Christoph was a steady presence throughout, always available and always supportive. I could not have asked for a better pair of advisors.

I am also grateful to Avi Feller and Julia Stoyanovich for joining my dissertation committee. Avi was consistently welcoming and generous during my visits to Berkeley, and I very much look forward to working together with Julia at NYU. I thank Jörg Drechsler for stepping in as examiner for the defense.

My PhD would have looked very different without Juan Perdomo. Juanky was something like a third advisor and a friend, and I learned an enormous amount from him, from how to give better talks and asking sharper questions to the importance of not taking it all too seriously. I deeply admire his relentlessness in finding the best version of a piece of work, and the genuine joy he brings to it. Our time together at Harvard is among my fondest memories of this PhD, and I very much look forward to continuing our work at NYU.

I have also been lucky to work with a wonderful set of coauthors and collaborators. I am especially grateful to Stefan Feuerriegel for his advice and insights into causality, and to Emily Aiken for bringing a development economics perspective to our work. Thank you also to Noam Barda and Julian Rodemann for being such generous collaborators at different stages of this PhD. And to Gesine Stefan, Max Kunaschk, and the team at the IAB, I am thankful for their patience in teaching me the intricacies of the BA.

I am also grateful to Cynthia Dwork for welcoming me for a research semester at Harvard, and to Sendhil Mullainathan for hosting me at MIT. Cynthia drew me into so many fascinating discussions and taught me a great deal about what theoretical work in fairness can look like. Sendhil's way of approaching difficult problems in the social domain with optimism and ambition has been a genuine inspiration. Thank you also to Moritz Hardt, whose research has been a real influence on this work, and who has always given me good book recommendations.

Thank you to everyone who made my time in the US so memorable: To Serena Wang for solving murder mysteries together; Chris Hays for the good conversations and always knowing the best places to eat; Benedikt Koch for facing election night together and always being there to answer questions and read drafts; Andreas Haupt for always turning up somewhere unexpected; and Carmel Baharav and Rachel Li for introducing me to the wonders of American cheesecake. I am also very grateful to Sílvia Casacuberta, who has become both a wonderful collaborator and a friend; working through calibration together continues to be a lot of fun, and I always enjoy our conversations.

I am also incredibly lucky to have spent my PhD at the SODA Lab, where I found not just colleagues but many good friends. Thank you to Sarah Ball for starting this journey together, all our adventures and keeping each other more or less sane, to Jan Simson for keeping me active on hikes, and to Felix Henninger and Patrick Schenk for always being people to count on. I am also so glad that so many wonderful people joined the lab along the way: Clara Strasser Ceballos for keeping the Spanish faction strong, Marcus Novotny for the great conversations and for showing me there is a world outside academia, and Lisa Schmierer for being a constant source of motivation and curiosity. I am thankful to Bolei Ma for the steady supply of candy and incredible research output, to Jacob Beck for being a great duo in getting DSSG off the ground, and to Olga Kononykhina for your optimism. Thanks go to everybody in the lab that joined and left over the years, to Caro Haensch and Malte Schierholz for being such good anchors of stability, and to Wiebke Weber, who always had a sympathetic ear for the moments when I was taking things a little too seriously.

I am also grateful to my friends at the Zuse School, and especially to Max Fleissner and Lukas Gosch for the many great evenings, hikes, and Padel attempts along the way. Thank you also to all the people I met through fencing, who provided an outlet for not fighting reviewers too directly. I am also thankful to Jonas Mikhaeil, a great friend and someone I have always felt intellectually connected to. It was a good idea, it turns out, to convince each other that statistics was worth pursuing. I am deeply grateful to Andreas Albiez for being a best friend for as long as I can remember – who always found a way to make my day. I thank Sofie Stoffel, for a friendship that has survived us constantly swapping continents, and Julian Zeller for many fun house parties, introducing me to all kinds of wines and beers, and for bringing me to the opera. Thanks to Viviane Nguyen for the many fun times together and finding the best food in Munich. I am also incredibly grateful to Sherry Föhr for her generous help with corrections and for her warmth and support throughout this journey.

Finally, I am deeply grateful to my family, my father Andreas, my mother Arantxa, and my sister Saioa, for their love and support throughout, and for keeping my artistic and creative sensibilities alive even as I wandered into statistics. I am also so incredibly grateful to Nanina Föhr, my partner, for her unwavering support, for being such a wonderful intellectual companion, and for making life joyful through all of it. None of this would have been the same without her.

Abstract

Public institutions increasingly rely on predictive algorithms to guide consequential decisions about citizens, from prioritizing job seekers for counseling to targeting social benefits. Yet many of these systems fail to deliver on their promise, and some provoke significant public backlash. A fundamental reason is that predictive systems are typically designed, evaluated, and deployed in isolation from the institutional and policy contexts they are meant to serve, resting on assumptions about data, decisions, and objectives that frequently conflict with the realities of public sector decision-making. Looking beyond the model itself is therefore a prerequisite for understanding when algorithmic systems create value for the populations and institutions they serve. We develop this perspective through two complementary lines of research, asking both whether a system's assumptions hold in context, and whether improving prediction is worthwhile relative to the other investments an institution could make.

We trace how distribution shifts, label bias, the influence of past decisions on training data, competing objectives, and human discretion create misalignments between machine learning models and public sector decision-making, and argue for reorienting modeling efforts from predictive accuracy toward decision-making outcomes. We show that this often requires moving to an estimand better aligned with the decision at hand, and carefully interrogating the causal assumptions connecting predictions to decisions that many predictive systems leave implicit.

At the same time, predictions are not an end in themselves but a means to improve downstream welfare and advance institutional objectives. This raises a more fundamental question about how valuable improving prediction is in solving allocation problems, relative to the other investments a planner could make. Indeed, we show through a stylized model and a large-scale case study on long-term unemployment in Germany that better predictions are often not the most important factor in improving outcomes. We then define a broader meta-design space of allocation problems and develop tools that help planners identify which components of an allocation matter most for downstream welfare. We demonstrate this approach in case studies on the targeting of cash transfers in Ethiopia and the profiling of job seekers in Germany.

Across these contributions, prediction emerges as one component within the broader design of institutional decision-making, one that must be analyzed in light of the societal context, operational constraints, and policy objectives it is meant to serve.

Zusammenfassung

Öffentliche Institutionen setzen zunehmend auf Vorhersagealgorithmen, um weitreichende Entscheidungen über Bürgerinnen und Bürger zu treffen – von der Prioritisierung von Arbeitsuchenden in der Beratung bis zur Vergabe von Sozialleistungen. Viele dieser Systeme bleiben jedoch hinter ihren Versprechen zurück und manche stoßen auf erheblichen gesellschaftlichen Widerstand. Ein zentraler Grund liegt darin, dass Vorhersagesysteme typischerweise losgelöst von den institutionellen und politischen Kontexten entwickelt, evaluiert und eingesetzt werden, denen sie dienen sollen. Sie beruhen auf Annahmen über Daten, Entscheidungen und Ziele, die mit der Realität öffentlicher Entscheidungsfindung oft in Konflikt geraten. Den Blick über das Modell hinaus zu richten ist daher unerlässlich, um zu verstehen, wann algorithmische Systeme tatsächlich einen Mehrwert für Bevölkerung und Institutionen schaffen. Diese Perspektive entwickeln wir entlang zweier komplementärer Forschungsstränge: Wir fragen zum einen, ob die Annahmen eines Systems im jeweiligen Kontext überhaupt gelten, und zum anderen, ob eine Verbesserung der Vorhersagen gegenüber anderen möglichen Investitionen einer Institution lohnenswert ist.

Wir zeigen, wie Verteilungsverschiebungen, Verzerrungen in den Zielgrößen, der Einfluss vergangener Entscheidungen auf die Trainingsdaten, konkurrierende Ziele und menschlicher Ermessensspielraum zu Diskrepanzen zwischen Modellen des maschinellen Lernens (ML) und öffentlicher Entscheidungsfindung führen. Wir plädieren dafür, den Modellierungsfokus von der Vorhersagegenauigkeit auf die Entscheidungsqualität zu verlagern. Dies erfordert häufig den Wechsel zu einer Zielgröße, die besser auf die jeweilige Entscheidungssituation zugeschnitten ist, sowie eine sorgfältige Auseinandersetzung mit den kausalen Annahmen, die Vorhersagen und Entscheidungen verknüpfen – Annahmen, die in vielen Systemen implizit bleiben.

Vorhersagen sind dabei kein Selbstzweck, sondern Mittel um das Wohlergehen der betroffenen Personen zu verbessern und institutionelle Ziele zu erreichen. Dies wirft eine grundlegendere Frage auf: Welchen Mehrwert bringt eine bessere Vorhersage für die Lösung von Allokationsproblemen – gemessen an dem, was alternative Investitionen leisten könnten? Anhand eines simplen Modells und einer groß angelegten Fallstudie zur Langzeitarbeitslosigkeit in Deutschland zeigen wir, dass bessere Vorhersagen häufig nicht der entscheidende Hebel für bessere Ergebnisse sind. Darauf aufbauend definieren wir ein breiteres Spektrum von Gestaltungsentscheidungen, die einem Allokationssystem vorgelagert sind, und entwickeln Werkzeuge, mit denen Planer erkennen können, welche dieser Entscheidungen den größten Einfluss auf das Wohlergehen der der betroffenen Personen haben. Wir illustrieren diesen Ansatz anhand von Fallstudien zum Armuts-Targeting in Äthiopien und zum Profiling von Arbeitsuchenden in Deutschland.

Insgesamt zeigt sich, dass Vorhersagen nur ein Baustein institutioneller Entscheidungsfindung unter vielen sind und sich nur vor dem Hintergrund des gesellschaftlichen Kontexts, operativer Rahmenbedingungen und der verfolgten Ziele angemessen beurteilen lassen.

Contents

Part I

Introduction

1 Introduction

> Quantification is a way of making decisions without seeming to decide.
>
> *(Theodore M. Porter, Trust in Numbers (1995))*

Algorithmic predictions have come to inform a wide range of high-stakes decisions across the public sector. In domains ranging from child welfare screening to fraud detection, predictive systems are used to determine where institutions direct their attention and who receives interventions. The appeal is considerable, particularly as public institutions generally operate in resource-constrained environments. Such systems offer the prospect of more efficient, more accurate, and fairer decision-making, applied consistently at scale. However, many systems have failed to deliver on that promise, producing unintended consequences, behaving in ways their designers did not anticipate, and attracting public backlash severe enough to force their withdrawal. These failures are often not failures of predictive accuracy in any narrow sense, but failures in the link between a model and the institutional setting it is embedded in.

We ask when, and under what conditions, predictive systems can actually provide value in public institutions. In pursuing this question, we look beyond the model itself to the institutional context in which it operates. This perspective is developed through two complementary lines of research.

The first (Chapters 4, 5) asks whether a system's technical assumptions actually hold in its deployment setting, and thus whether the system is doing what it claims to do. When the assumptions connecting data, predictions, and decisions fail to match the deployment context, even a seemingly accurate model can be misaligned with the policy goal it is meant to serve. The second (Chapters 6, 7) asks whether improving prediction is worth it at all, given that a model is only one component of a larger, resource-constrained system and should be weighed against other investments a planner could make instead. Here we largely take the institution's stated objective as given and the prediction problem as well-posed, setting aside the complications of the first line in order to show that the value of prediction is contextual even in a setting otherwise favorable to it.

What emerges from both lines of work is an understanding of evaluation that requires making explicit what a narrow focus on the model alone renders invisible — the assump-

tions on which it rests, and the alternatives against which it should be judged. The value of a predictive system is, we argue, necessarily relative to the institutional context it is embedded in.

Chapter 2 acts as an introduction to the dissertation by providing background on algorithmic systems in the public sector, introducing the ideas this work builds on, situating it within the broader literature, and locating predictive systems within a longer history of quantification in public institutions. Chapter 3 then introduces the four contributing articles, before each is reproduced in Chapters 4–7. Chapter 8 closes with an outlook on open problems.

2 Prediction in the Public Sector

This chapter provides background for the contributions of this dissertation, with emphasis throughout on how predictive systems are embedded in public institutions and the technical and institutional assumptions underlying their deployment. Section 2.1 takes stock of the algorithmic systems now in use across the public sector and scopes what prediction systems we focus on. Section 2.2 turns to their mixed track record, using two prominent cases to argue that failures often stem from a mismatch between a model's technical setup and the institutional context. Section 2.3 zooms in on this link, formalizing common modeling frameworks and showing how choosing the appropriate one depends on assumptions about how predictions, outcomes, and decisions relate. Section 2.4 brings in the broader scholarship on the tensions between algorithmic decision-making and public institutions, and situates these systems within a longer history of quantification in government. Finally, Section 2.5 returns to the allocation problem this dissertation focuses on, asking what role prediction plays within it.

2.1 Taking Inventory

Machine learning algorithms have proliferated throughout public institutions. Engstrom et al. (2020) find that close to half of all US federal agencies have begun to use and experiment with machine learning systems, spanning policy domains from financial regulation to social welfare and functions from internal administration to frontline service delivery. van Noordt and Misuraca (2022) document a similarly broad spectrum of applications across the European Union.

While these surveys already point to algorithmic systems being widely embedded in government, recent years have seen a marked acceleration. In German federal administration, for instance, the government reported 212 distinct AI applications in 2024, up from 113 the year before (Bundesregierung 2024). The *Marktplatz der KI-Möglichkeiten*, a central register of AI systems in German public administration launched in 2026, records a near-exponential rise in registered systems over recent years (Bundesministerium für Digitales und Staatsmodernisierung 2026). Part of the apparent growth also reflects improved and more systematic reporting, but it seems undeniable that there is an administrative

zeitgeist in which adopting AI has increasingly become an institutional and political priority.

Forming an accurate picture of this landscape is nonetheless difficult, in part because there is little agreement on what should count as a relevant system to begin with. The New York City Automated Decision Systems Task Force, established in 2018, was unable to reach consensus on a working definition of an automated decision system, including whether the term should extend to tools such as spreadsheets or internet searches (NYC ADS Task Force 2019). Reporting is often intransparent or incomplete, and frequently occurs on a voluntary basis (Vieth-Ditlmann et al. 2025). There is also a sense that definitional ambiguity is sometimes put to strategic use by institutions, downplaying or inflating the role of AI within an institution as the occasion demands.

In practice, AI, machine learning, algorithmic decision-making and predictive system are used more or less interchangeably to describe a range of computational systems[1]. Among these systems, this work focuses on predictive systems, algorithms that estimate decision-relevant outcomes from data in order to inform individual-level decision-making. Such systems differ from rule-based or expert systems, which apply manually specified decision rules. The output of predictive systems may then be used to automate the decision outright, or it may serve as a signal to a human decision-maker. Within this class, a common distinction is made between systems trained to replicate past human decisions (Barocas et al. 2023), like, a model scoring job applicants on historical hiring patterns, and systems targeting (social) outcomes directly relevant to the decision, such as the probability of a health complication or of finding employment within a given horizon. The distinction can blur in practice, because social outcomes are typically only observable through the lens of past institutional responses.

Borrowing in spirit from the EU AI Act (*Regulation (EU) 2024/1689 (Artificial Intelligence Act)* 2024), we focus on predictive systems that inform high-stakes decisions about whether individuals receive a particular intervention, service, or resource, in domains such as health, welfare, and education (i.e., algorithms for "care" (Brayne 2017; Johnson and Zhang 2022)). The primary setting is the public sector, though much of the discussion carries over to sensitive commercial domains such as lending or insurance. What falls outside this study's scope are systems whose individual decisions carry comparatively low-stakes, such as recommendation, advertising, or content ranking, even where their aggregate effects on public life may be significant, as well as systems performing narrow technical tasks such as OCR, speech recognition, or document classification, even where these may be embedded in consequential pipelines.

[1] For example, this OECD report on AI in the employment service (Brioscú et al. 2024) is slightly confused in its taxonomy and argues that "AI-enabled profiling is usually referred to as statistical profiling, albeit most often, the underlying model is not a statistical model but a much more sophisticated prediction model."

2.2 Contested Algorithms

The track record of predictive algorithms in the public sector has been mixed. Such systems are often promoted as a way to improve the efficiency of existing operations, enhance service delivery, and support more "objective", transparent, and standardized decision-making (Engstrom et al. 2020; Saxena et al. 2021; Barocas et al. 2023). These hopes are particularly salient in the public sector, where institutions operate under tight resource constraints and where human decision-makers exercise considerable discretion – discretion that such tools promise to "discipline" and align with institutional goals (Lipsky 2010). The German Federal Employment Agency (*Bundesagentur für Arbeit*, BA), for example, expects to lose over 40% of its workforce by 2034 and hopes that automation and the adoption of AI and ML can help relieve this demographic pressure (Bundesagentur für Arbeit 2024). Ludwig et al. (2024) offer an optimistic account of using algorithms to address public policy problems. They argue that many such algorithmic interventions create net benefits for society while reducing costs, giving such programs an effectively infinite marginal value of public funds, a "free lunch" for government. This potential stems in part from the weaknesses of the human decision-making these tools replace and from the ability to cheaply scale up effective algorithms. At a minimum, they argue, there exists a subset of highly impactful algorithms that offer a genuinely new solution concept for many policy problems.

However compelling these accounts, such systems are also deployed in settings where the costs of getting decisions wrong are especially high. Public institutions operate at large scale, often under ambiguous mandates, and frequently serve vulnerable populations who depend on them as their only available provider. In recent years, the harms and failures associated with algorithmic decision-making have been documented across a wide range of domains, including unemployment fraud (Charette 2018), social welfare (Eubanks 2018; Guo et al. 2025), employment services (Achterhold et al. 2025), criminal justice (Angwin et al. 2016; Pruss 2023), child welfare (Das 2022), and health (Obermeyer et al. 2019).

In many of these cases, deployed systems have failed to deliver on their stated objectives, or at least failed to convince key stakeholders that they do. Systems have frequently been contested by affected populations, regulators, institutional stakeholders, and the wider public, and abandoned after years of development and substantial institutional investment (Johnson et al. 2024). The following two prominent cases are instructive.

AMS Algorithm Beginning in 2015, the Austrian Public Employment Service (*Arbeitsmarktservice*, AMS) developed an algorithmic system to assist in the allocation of support to unemployed jobseekers (Allhutter, Cech, et al. 2020; Allhutter, Mager, et al. 2020). The system, known as the *AMS Algorithm*, was trained on historical administrative data to estimate the probability of individual jobseekers finding employment within a de-

fined time horizon. Based on these estimates, individuals were segmented into three risk groups, which in turn structured the type and intensity of support they would receive.

The plan was to focus institutional attention on those assigned to the medium-risk bucket. Following a logic of assumed cost-effectiveness, individuals at low risk were not seen as in need of support, while interventions for high-risk individuals were considered not to be efficient (Allhutter, Cech, et al. 2020).

The system attracted significant public and academic criticism. Concerns ranged from the use of sensitive attributes that created the risk of reinforcing existing disadvantages to the seemingly ad hoc and opaque way the system formalized the agency's objectives and operationalized its predictions. For example, why did the system operate on three risk buckets and assume homogeneous treatment within each? In August 2020 the system was prohibited by Austria's Data Protection Authority. The legal dispute was settled in 2025 (Eurofound 2025) in favour of the employment offices. However, AMS had been forced to delete all project-related data while the case was pending, rendering future use of the system unlikely (ORF 2025).

Toeslagenaffaire The Dutch childcare benefits scandal (*toeslagenaffaire*) involved false allegations of welfare fraud against close to 30,000 parents. It led to the resignation of the cabinet of Prime Minister Mark Rutte in January 2021 after the release of a parliamentary report pointedly titled *Unprecedented Injustice* (Hadwick and Lan 2021).

The Dutch tax administration deployed several algorithmic tools to process benefit applications at scale and detect fraud. For example, it included a classifier trained on past cases that flagged applications for a manual audit. In isolation, the design of the algorithm was not particularly noteworthy. What mattered was how it was embedded in the institution, both in how its training data was produced and how its predictions were used to make decisions.

On the input side, tax officials exercised considerable discretion about which cases to audit. They often used the digital infrastructure to run manual queries targeting specific nationality groups perceived to be more prone to fraud (Hadwick and Lan 2021). The outcomes of these audits fed back into the model's training data, creating a self-reinforcing loop of discriminatory selection practices.

On the output side, the Dutch tax administration adopted an "all or nothing" approach. Even minor mistakes in paperwork forced retroactive repayment of benefits received over years, leading many families into financial ruin. The design of the algorithm was not responsible for this policy. However, by increasing the number of cases audited, it exposed far more families to its consequences (Hadwick and Lan 2021). Only a small subset of caseworkers had direct access to the model, making it difficult for affected families to learn why they had been flagged, let alone contest it.

The system operated with human decision-makers at multiple points in the pipeline. However, this did not lead to effective oversight, as the algorithmic tools were used to

scale up problematic practices already present in the institution.

In both cases, the failure lies not solely in the model's predictive accuracy but in the assumptions connecting it to its institutional setting: which outcome is predicted and to what end, how the available data was generated, and how scores are turned into decisions. What matters, then, is which implicit assumptions a given modeling choice carries, and whether they are plausible in the context where the system is used.

2.3 From Predictions to Decisions

A natural starting point for making these implicit assumptions explicit is to formalize the modeling frameworks commonly used in the public sector, asking for each what the system estimates and when that estimand can plausibly inform the decision an institution wants to make. Whether it does depends on what the institution is trying to achieve in the first place, and on external assumptions, informed by institutional context, about how predictions, outcomes, and decisions relate. The following discussion aims to be applicable across a range of model classes, with the focus on the relevant estimand and its connection to the downstream decision-making. See Fischer-Abaigar et al. (2024) and Liu et al. (2026) for a similar discussion of prediction and decision-making.

2.3.1 Risk Prediction

Prediction in the public sector often follows a simple recipe (Wang et al. 2023; Liu et al. 2026). First, train a predictive model of some outcome believed to be relevant for the decision at hand on historical data. Second, use the resulting predictions to prioritize, triage, or screen individuals for downstream interventions.

More formally, let $X \in \mathcal{X}$ denote the observed characteristics of an individual, such as demographic attributes, administrative records, or case histories, and let $Y \in \mathcal{Y}$ denote an outcome of interest. Common prediction targets include the probability of an adverse event, such as a health complication (Obermeyer et al. 2019), child maltreatment (Coston, Mishler, et al. 2020), or long-term unemployment (Kern et al. 2024). Prediction then follows standard supervised learning. Given an observed dataset $\{(X_i, Y_i)\}_{i=1}^n$ drawn from some distribution over pairs $(X, Y) \sim \mathcal{D}$, the goal is to learn a predictor approximating the conditional expectation

$$f(x) \approx \mathbb{E}_{\mathcal{D}}[Y \mid X = x]$$

Predictions are not an end in themselves, but typically used to inform institutional decision-making process. For ease of presentation, let $A \in \{0, 1\} \subseteq \mathcal{A}$ denote a binary action taken for an individual, such as whether to allocate a medical intervention, enroll someone in a job training program, or flag a case for further review. Following classical sta-

tistical decision theory (Berger 1985), a simple way to connect predictions to decisions is through a utility function,

$$u\colon \mathcal{Y} \times \{0,1\} \to \mathbb{R}$$

where $u(y,a)$ specifies the value of taking action a when the relevant outcome is y. We can then evaluate a decision policy $\pi\colon \mathcal{X} \to \{0,1\}$ through its expected utility,

$$V(\pi) := \mathbb{E}[u(Y, \pi(X))]$$

For example, an institution seeking to prioritize those most in need might value acting on individuals whose outcome exceeds a threshold t, $u(y,a) = a \cdot (y - t)$. The policy maximizing V then acts whenever the expected outcome is high enough,

$$\pi(x) = \mathbf{1}\{\mathbb{E}[Y \mid X = x] > t\}$$

In practice, of course, the conditional expectation $\mathbb{E}[Y \mid X]$ is unknown, so a common strategy[2] is to estimate it from data $f(x) \approx \mathbb{E}[Y \mid X = x]$ and respond to the estimate. A simple predict-then-optimize strategy then amounts to ranking individuals by predicted risk $f(x)$ and targeting those at the top

$$\pi(x) = \mathbf{1}\{f(x) \geq t\}$$

Systems built on this logic are variously referred to as risk prediction, risk assessment tools (Coston, Mishler, et al. 2020) or predictive optimization (Wang et al. 2023), reflecting that they typically predict adverse events to prioritize downstream support.

In practice, the utility being pursued is often left implicit or difficult to fully formalize. Nevertheless, treating it as an explicit object is useful for interrogating the assumptions underlying the connection between predictions and decisions. One important distinction is whether the outcomes Y entering the utility depend on the decision A, or whether the relevant state of the world is independent of the decision taken. Kleinberg et al. (2015) call problems of the latter kind *prediction policy problems*, since a prediction model is sufficient to inform the decision without explicit causal inference. For example, in medical diagnostics, the presence of a disease does not depend on the decision to treat (Alur et al. 2024); in fraud detection, whether fraud occurred does not depend on whether we choose to investigate.

Example 1. The example Kleinberg et al. (2015) use in their seminal paper is joint replacement surgery for patients suffering from osteoarthritis. They argue that patients who die

[2] The plug-in procedure is common practice, even if not necessarily optimal for a given downstream objective. Later, we will discuss approaches that directly learn a utility maximizing decision policy $\pi\colon \mathcal{X} \to \mathcal{A}$ without separating learning from decision-making (see Section 2.5.1).

within a year after surgery do not live long enough for it to have been beneficial, posing a prediction policy problem with a simple payoff structure: predict mortality risk, and exclude those most at risk in order to redirect resources toward lower-risk patients.

Framing it as a purely predictive problem involves a slight sleight of hand, since the utility itself encodes prior causal and domain knowledge about how we expect predictions to relate to downstream welfare. In Example 1, we need to assume that surgery does not significantly affect mortality — Kleinberg et al. (2015) operationalize this by excluding deaths in the first month to separate baseline mortality from postoperative complications. We also need to find a meaningful mortality threshold above which surgery is no longer beneficial, and that the benefit of surgery is sufficiently homogeneous conditional on risk to justify ranking patients by predicted mortality alone. In effect, framing a problem as purely predictive does not eliminate causal assumptions but relocates them from the estimation step to the initial formulation of the decision problem.

2.3.2 Counterfactual Prediction

In many settings, however, the outcome that would inform the best decision is also the outcome the decision itself affects. In the language of potential outcomes (Rubin 1974; Splawa-Neyman et al. 1990), the outcome of an individual now depends on the action A taken. Let $Y(a)$ denote the potential outcome under action a. For each individual we observe only the outcome corresponding to the realized action, $Y = Y(A)$, while the counterfactual outcomes under the actions not taken remain unobserved.[3] For example, we may be interested in estimating the baseline risk,

$$\mu_0(x) := \mathbb{E}[Y(0) \mid X = x]$$

the expected outcome in the absence of an intervention. This is a common target in public sector applications, when the goal is to identify individuals most in need of support.

Example 2. Coston, Mishler, et al. (2020) discuss the screening of child welfare hotline calls in Allegheny County. The goal is to identify the children most at risk of maltreatment if not investigated (i.e., $Y(0)$), in order to decide which calls to follow-up. Whether an investigation occurs will likely affect downstream risk, so the outcome of interest is causally connected to the decision being informed.

Recent work refers to such tasks as *counterfactual prediction*, because the goal is to predict what would happen under a specified action (Coston, Kennedy, et al. 2020; Coston,

[3] Writing $Y = Y(A)$ presumes the conditions usually collected under the Stable Unit Treatment Value Assumption (SUTVA), namely that the observed outcome equals the potential outcome under the realized action (consistency), and that the outcome of an individual is not affected by the actions assigned to others (no interference).

Mishler, et al. 2020; Guerdan et al. 2023). The boundary with what Kleinberg et al. (2015) call prediction policy problems can be somewhat ambiguous. For example, they mention predicting unemployment spell length to inform job-search strategies as a prediction policy problem, although such strategies may themselves affect the unemployment spell being predicted. Similarly, in bail prediction, the relevant outcome is observed only for defendants who were released (Kleinberg et al. 2018; Rambachan et al. 2022). While these are still framed as pure prediction tasks, they differ from ordinary supervised learning in that the historical record only reveals outcomes under the decisions that were actually made.

2.3.3 Causal Inference

In many problem settings, we may wish to model the expected benefit of an intervention directly (Athey 2017). That is, rather than directing resources toward individuals with the worst expected outcomes, an institution may focus on those for whom the intervention improves outcomes the most. The conditional average treatment effect (CATE),

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

is a common estimand in causal inference and captures how much we would expect an individual with covariates x to benefit from receiving treatment.

Example 3. Cockx et al. (2023) and Goller et al. (2025) study the heterogeneous effects of job training and assistance programs for jobseekers. Their primary interest is the effect of the program on reducing unemployment duration, and whether heterogeneity in this effect across individual characteristics x, for example by educational background, can be exploited to better allocate program slots.

Identification Neither the counterfactual prediction target $\mu_a(x)$ nor the CATE $\tau(x)$ can be estimated directly from observational data without additional assumptions. If we were to estimate these causal quantities with simple averages of the observed outcomes, we would obtain estimates that reflect the rules of past institutional decision-making as much as the underlying outcome process. For instance, if individuals at higher risk were more likely to receive the intervention, then the labels observed will be disproportionately drawn from lower-risk cases. A model trained naively on these labels may consequently underestimate baseline risk precisely where it matters most.

Much can be said about the challenge of *identification*, linking a causal estimand to a statistical quantity that can be estimated from observed data; see Imbens and Rubin (2015) and Hernán (2023) for a detailed discussion. For completeness, we note that many relevant estimands can be identified under *strong ignorability*:

Assumption 2.3.1 (Strong Ignorability)**.** *The following two conditions hold:*

(i) Unconfoundedness. *Conditional on observed covariates, treatment assignment is independent of the potential outcomes,*

$$Y(1), Y(0) \perp A \mid X$$

(ii) Overlap. *For all x in the support of X,*

$$0 < \Pr(A = 1 \mid X = x) < 1.$$

Randomized experiments satisfy these assumptions by design, but are often infeasible in public sector settings due to ethical and legal constraints. In New Zealand, a proposed study of a child-maltreatment risk model was halted because it would have required generating risk scores for newborns without using those scores to trigger services during the study period, effectively waiting to see whether adverse outcomes would materialize for flagged children[4] (Jones 2015). In observational settings, unconfoundedness in particular is difficult to verify, especially when decisions were informed by factors not recorded in the data, such as the first impression a job seeker made during a caseworker interview.

When they do hold, they imply inter alia that

$$\mu_a(x) := \mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y \mid X = x, A = a]$$

so that $\mu_a(x)$ and the conditional average treatment effect $\tau(x)$ are identified from observational data. This allows us to evaluate the expected utility

$$V(\pi) = \mathbb{E}[u(Y(0), Y(1), \pi(X))]$$

of a proposed decision policy π, provided the utility decomposes into identified quantities. For instance, a planner who values the outcome realized under the policy

$$u(y(0), y(1), a) = a \cdot y(1) + (1 - a) \cdot y(0)$$

obtains $\mathbb{E}[\pi(X)\mu_1(X) + (1 - \pi(X))\mu_0(X)]$. However, this is not true for arbitrary counterfactual utilities that depend jointly on $y(0)$ and $y(1)$. For example, if we wish to allocate resources only to those who respond to treatment, the corresponding utility,

$$u(y(0), y(1), a) = a \cdot \mathbf{1}\{y(1) > y(0)\}$$

is not identified under strong ignorability, because it depends on the joint distribution of $(Y(0), Y(1))$. Koch and Imai (2025) formalize this distinction by showing that counter-

[4] The minister responsible reportedly noted "these are children, not lab rats" in the margins of the proposal (Jones 2015).

factual utilities are identified under strong ignorability if and only if they are additive in the potential outcomes.

2.3.4 From Institutions to Estimands

As we have seen, different problem formulations lead to different estimands and different identification requirements, and imply different connections between predictions and decisions. But which modeling setup is appropriate in a given setting is not something the statistical framework can answer on its own. It requires an understanding of the institutional context, including what a decision-maker is trying to achieve, how actions are expected to affect outcomes, which outcomes matter, and how historical labels were generated. Otherwise, implicit or poorly justified assumptions can be carried through the pipeline as if they were merely technical modeling choices.

Employment services provide a useful example. Many public employment agencies use profiling systems to predict the risk of long-term unemployment (Desiere and Struyven 2021; Kern et al. 2024), and existing evaluations focus largely on predictive performance (Gallagher and Griffin 2023). But why an accurate prediction of long-term unemployment should be the basis for allocating support is a separate question, one that depends on an institutional rationale that is often left ambiguous. In practice, profiling systems are justified both on grounds of cost-effectiveness – directing resources where they are expected to have the largest impact, and on normative grounds – identifying those most in need of support (Desiere and Struyven 2021). Whether a given risk score is meant to serve one goal or the other is not always spelled out, but the two can require very different assumptions and modeling approaches.

When a planner does commit to a specific objective, the assumptions connecting predictions to that objective need to be interrogated carefully. The Austrian AMS algorithm, discussed earlier, directed resources toward individuals in a medium-risk category on the assumption that this group would benefit most from intervention (Allhutter, Cech, et al. 2020). The predictive model itself was a straightforward instantiation of the risk prediction framework described above. But the assumption connecting predictions to the allocation rule was an implicit causal claim about treatment effect heterogeneity, and newer research finds no clear evidence that medium-risk job seekers actually benefit more from support than those at high risk (Eppel et al. 2026). This kind of brittleness is not visible when evaluating the predictive model in isolation, and becomes apparent only when considering how and why predictions are embedded into institutional decision-making.

Being explicit about objectives can also help make informed tradeoffs when choosing an allocation strategy. The estimand most directly aligned with the decision objective may be difficult to estimate reliably, while a simpler prediction target may be less directly aligned but more stable, easier to validate, or less demanding in its assumptions. Recent work studies this type of tradeoff, asking when simpler targeting rules based on baseline risk can perform well relative to rules based on heterogeneous treatment effects

(Fernandez-Loria and Provost 2022a; Athey et al. 2025; Sharma and Wilder 2025). In this sense, the relevant question is not only which estimand is formally closest to the objective, but which target supports the best decision under the available data and assumptions.

2.4 Algorithmic Decision-Making in Context

A substantial body of literature investigates the challenges of integrating predictive systems into institutional decision-making. Many disciplines have engaged with the legal, regulatory, institutional, and technical challenges of setting up, governing, contesting, auditing, procuring, and maintaining such systems in the public sector. To name only a few entries in this broad field, Wirtz et al. (2019) and Levy et al. (2021) provide overviews of the open problems raised by algorithmic decision-making and AI systems in the public sector. With a more specialized focus on legal questions, Richardson (2021) develop legal definitions for understanding automated decision-making in public institutions, while Engstrom and Haim (2023) examine the regulatory tradeoffs between encouraging innovation and ensuring adequate oversight. A separate body of work studies how the advent of algorithmic decision-making reshapes power relations, legitimacy, and due process inside public institutions (Waldman 2019; Lazar 2024). Other work highlights the institutional challenges AI poses for public administration (Agarwal 2018), including in the U.S. federal government (Engstrom et al. 2020) and across the European Union (van Noordt and Misuraca 2022). Sun and Medaglia (2019) and Saxena et al. (2021) map out the operational challenges of integrating such systems from the perspective of frontline caseworkers and other stakeholders.

2.4.1 Algorithmic Fairness

On the technical side, *algorithmic fairness* has been the dominant lens through which the computer science and statistics community has examined the failures and harms of algorithmic decision-making (Chouldechova and Roth 2020; Barocas et al. 2023). We cannot do justice to such a sprawling field here, but point to Mehrabi et al. (2021) and Mitchell et al. (2021) for a more comprehensive treatment. Broadly speaking, the field asks what makes an algorithmic system unfair and develops tools to measure and mitigate disparities in decision-making and prediction across individuals and salient subgroups.

Such disparities may, for example, stem from pre-existing biases in the data[5] (Barocas and Selbst 2016), from the design of the algorithm itself, or from how its predictions are used in downstream decisions. Because predictive systems in the public sector make de-

[5] What constitutes bias is a subtle question (Friedman and Nissenbaum 1996) and will always involve some level of value judgment about what type of inference we hope to draw from the data. For example, we may be concerned about having undersampled specific groups, or about propagating historical patterns through a decision process without due consideration.

cisions with particularly high stakes for affected individuals (Eubanks 2018), much of the development of algorithmic fairness has proceeded hand in hand with the documentation and analysis of failures in this setting.

Considerable work in the field has been devoted to proposing formal definitions of what it means for an algorithm to be fair (Chouldechova and Roth 2020). This has produced a battery of fairness criteria, often in tension with one another[6] or other policy objectives (Kleinberg et al. 2016; Chouldechova 2017; Corbett-Davies et al. 2017), ranging from individual-level similarity metrics (Dwork et al. 2012) and counterfactual fairness notions (Kusner et al. 2017) to group-level measures such as equalized odds (Hardt et al. 2016) or (subgroup) calibration (Hebert-Johnson et al. 2018; Mehrabi et al. 2021; Mitchell et al. 2021). While much of this work has remained centred on supervised classification, fairness criteria and methods have been developed for many other settings, such as ranking (Yang and Stoyanovich 2017) and image generation (Laszkiewicz et al. 2024).

Given some formal definition of fairness, a large range of methods exists to ensure that a system satisfies a chosen fairness metric, whether by transforming the input data, adjusting the learning procedure, or modifying the decision rule after training, for example by setting group-specific thresholds in a classifier (Mehrabi et al. 2021). The sheer variety of available metrics and methods already points to the fact that evaluating the fairness of a system cannot be a purely technical enterprise, but requires normative and contextual input (Saleiro et al. 2018; Pechenizkiy et al. 2025). In fact, the field has faced criticism for focusing too narrowly on method development in stylized settings (Laufer et al. 2022).

One immediate way in which this broader context manifests itself is through the data. Fairness evaluations depend not only on the choice of metric but on a range of upstream decisions surrounding data collection, preprocessing, and evaluation design that are rarely made explicit (Stoyanovich et al. 2022; Simson, Fabris, et al. 2024). Recent work shows that seemingly innocuous choices, such as the imputation strategy for missing values, or how variables are transformed, can substantially alter fairness conclusions even when the same dataset and metrics are used (Caton et al. 2022; Guha et al. 2024; Simson, Fabris, et al. 2024).

Similarly, how a problem is formulated in the first place will shape any downstream fairness evaluation. At a basic level, even the choice of a decision threshold, often reflecting institutional constraints, can already determine which disparities become visible and which groups are most affected (Kern et al. 2024; Simson, Pfisterer, et al. 2024). More broadly, the process by which policy goals are operationalized into prediction targets involves choices that are often poorly documented and shaped by negotiation among stakeholders, and can have non-obvious downstream (fairness) consequences (Passi and Barocas 2019; Mitchell et al. 2021).

More generally, fairness evaluations rest on a range of assumptions about the data, the prediction target, normative goals, the relevant groups, and the link between model output and downstream decisions, few of which a made explicit (Mitchell et al. 2021). Selbst et

[6] Though less so in practice than classical results suggest (Bell et al. 2023).

al. (2019) warn of the dangers of abstracting away the socio-technical context in which systems operate, effectively treating fairness as a property of a model rather than of a system embedded in an institutional process. Pechenizkiy et al. (2025) make a related argument, noting that much of the fairness literature has overly focused on limited benchmarks while the broader lifecycle in which systems are designed, deployed, and maintained is often overlooked; they call for fairness evaluations to integrate an extrinsic assessment of how a system actually functions in its institutional setting.

2.4.2 When Predictive Systems Fail on Their Own Terms

Work on algorithmic fairness has shown that to evaluate whether a system is fair, choices about data, prediction targets, and how predictions connect to decisions need to be examined in their institutional context. However, these same choices also shape how well the system performs the institutional function it claims to serve in the first place. If this has never been plausibly established, claims about what the system achieves – fairness included – become difficult to interpret or defend. A growing line of work takes up this question directly: whether a system actually does what it claims is treated, as a baseline requirement for its legitimate use, even under a narrow reading of the institution's own objectives[7].

Coston et al. (2023) frame this as *validity*, the requirement that a "system [does] what it purports to do", which they argue is a prerequisite for evaluating values such as fairness and transparency. Raji et al. (2022) hold that it should be established whether an AI system functions at all before its broader (ethical) implications are debated; and Wang et al. (2023) treat basic functioning as a minimum condition for the legitimacy of *predictive optimization*[8], one they argue is frequently not met for this class of systems.

Returning to the AMS algorithm, it is questionable whether the prediction target chosen actually aligns with the institution's stated goal of improving the cost-effectiveness of their program allocation. Even accepting that target as given, recent research has found the deployed model prone to substantial misclassification, with error rates above one third on a dataset of young jobseekers (Achterhold et al. 2025). Either concern calls the system's efficacy into question before even considering further issues, such as whether it reinforces structural inequities in the labor market.

A common thread across these critiques, and one this dissertation develops further, is a shift in focus from a system's predictive accuracy to the assumptions that connect

[7] This does not imply that institutional objectives are singular, stable, benevolent, or just; agencies often pursue multiple and conflicting goals, and a system justified in terms of efficiency may also be entangled with aims such as expanding private digital industries (Collington 2022), legitimizing cost-cutting, or increasing administrative control. Fairness, too, can be part of an institution's stated objective, so this question does not always separate cleanly from the concerns above. The narrow question here is whether a system can plausibly perform the concrete claim put forward to justify its deployment.

[8] The class of systems they consider is similar to those discussed in Section 2.3.1 and to Kleinberg et al. (2015)'s notion of *prediction policy problems*.

its formalization to the institutional context it operates in. Fischer-Abaigar et al. (2024) characterize common misalignments between machine learning models and public sector decision-making and argue for making underlying assumptions explicit, and for shifting modeling efforts toward estimands that better match the structure of the decision problem at hand (Chapter 4). Liu et al. (2026) extend this agenda, arguing that predictive systems should be understood as policy interventions in complex social systems.

2.4.3 Formalization in Street-Level Bureaucracies

The preceding sections have argued that evaluating predictive systems requires examination of the assumptions that connect their formalization to the institutional context in which they operate. The challenge of formalizing complex decisions in public institutions is not new. Efforts to introduce mechanical decision rules, standardized procedures, and quantitative tools into government have a long history, and the tensions they produce between expert judgment and standardization recur across decades of public administration.

We have neither the ambition nor the historiographic authority to provide a comprehensive historical account, but see value in tracing a few threads that help locate current algorithmic efforts within a longer tradition of quantification in the public sector. The social and political processes that have historically driven the adoption of formal decision tools in government, the recurring conflicts they have produced, and the limitations long recognized in formalizing public decisions all shape the institutional reality into which algorithmic systems are now introduced. Appreciating this history helps both to recognize what is not unique to (predictive) algorithms in the difficulties and contestation surrounding their deployment, and to sharpen what, if anything, may be genuinely new.

We begin with a short discussion of the origins of quantification in public decision-making. We then zoom in on street-level bureaucracies, the institutional sites where many algorithmic tools are deployed, and the tensions between formalization and frontline discretion that have long characterized them. Finally, we turn to the policy implementation literature, which has studied the gap between the design of well-intentioned administrative interventions and the reality of their execution, and close by asking what may distinguish predictive systems from earlier formalization efforts.

Quantification of Public Decisions Porter (1995) provides a canonical account of how, over the course of the 19th and 20th centuries, decision-making in public institutions became increasingly dominated by quantitative expertise[9]. His central claim is that this should not be understood as a natural development of professional communities building ever more powerful methodologies, but seen through the lens of legitimacy

[9] Porter (1995) does not focus exclusively on public decision-making, but describes a range of professional communities and sciences. However, he frequently describes public bureaucracies as particularly susceptible to the political and social forces he describes.

and trust. Public bureaucrats who lacked an explicit democratic mandate, and whose expert judgment was increasingly subject to outside scrutiny, came under political and social pressure to justify the decisions they were making. Quantification and standardized decision-making gained ground as a way to constrain professional discretion and defend public decisions against charges of arbitrariness and bias. Porter calls this "mechanical objectivity", the ideal of explicit, rules-bound decision-making that minimizes individual judgment.[10] As he puts it, "quantification is a way of making decisions without seeming to decide" (Porter 1995, p. 7).

Porter traces this dynamic across a range of professional communities and institutional settings. He finds that professional communities that lacked social capital or cohesive elite networks were least able to defend their discretionary authority against outside demands for transparency and standardization. In countries where bureaucratic elites were less trusted, the push toward impersonal, rule-bound decision-making was correspondingly stronger.

For example, cost-benefit analysis emerged as a regulated and standardized practice within the U.S. Army Corps of Engineers, driven in large part by political pressure from Congress, lobby groups, and rival governmental agencies (Porter 1995, p. 148). Initially a fairly local and discretionary practice among engineers analyzing water infrastructure (Porter 1995, p. 186), cost-benefit analysis in U.S. federal agencies proliferated under the guidance of economists during the 1950s and 1960s, being applied to measure benefits in education, healthcare, welfare and the value of landscapes. This trend was encouraged by the development of operations research and systems analysis in the intellectual environment of the RAND Corporation (Harcourt 2018). Under Robert McNamara, systems analysis was adopted for military procurement and then expanded to governmental decision-making more broadly during the Johnson administration, emblematic of what Lemov et al. (2019) call a "cold-war rationality" that preferred mechanical, algorithmic approaches to public decisions over subjective professional judgment.

These efforts at expanding mechanical decision-making into public administration soon attracted criticism. Hammond (1966) argued that benefit-cost analysis [sic] had been "exalted into a 'secular sacrament'" while the decisions underlying such analyses were often poorly specified, resting on non-technical choices whose problem was "not their presence, but the fact that they may be tacit, inconsistent, or ill-chosen" (Hammond 1966, pp. 196, 208). Even proponents of expanding systems analysis into civilian government, such as Edward Quade of the RAND Corporation, warned of being so "mesmerized by the beauty and precision of the numbers that we overlook the simplifications required to achieve that precision," and cautioned that "big decisions, therefore, cannot be the con-

[10] It is important to distinguish between this notion and the general use of quantitative reasoning. The point is not the application of mathematical methods as such, but the adoption of standardized quantitative procedures where the process itself, rather than the expertise of the person applying it, is meant to guarantee the result (Espeland 1997).

sequence of a computer program or any application of a mathematical model" (Quade 1966, pp. 26, 28).

It is difficult to overlook the parallels to current debates around algorithmic decision-making, about what tools actually measure, what design decisions are formalized and what is implicitly left out, and whether mechanized decision rules make public decisions more accountable or less. We may call for making the design decisions underlying algorithmic systems more explicit, but it is arguably their seemingly technical, non-political nature that often makes them attractive to public institutions in the first place.

The tension between formalization and professional discretion that runs through the history of quantification[11] does not disappear once mechanical rules are in place. Even standardized procedures typically leave considerable room for how they are applied (Espeland 1997). To understand how this tension plays out in practice, it is necessary to look more closely at the institutions in which the algorithmic tools discussed in this dissertation are embedded.

Street Level Bureaucracies Street-level bureaucracies are public agencies characterized by significant day-to-day interactions with the public, staffed by front-line workers, such as caseworkers in employment offices, social workers, teachers, or police officers, who exercise considerable discretion in the decisions they make about service delivery. The concept was introduced by Lipsky (2010) in his eponymous book, originally published in 1980. These institutions are of particular interest here because they are the sites where many of the predictive tools discussed in this dissertation are deployed, and understanding their structural characteristics helps explain the difficulties that algorithmic tools face in these settings.

In street-level bureaucracies, front-line workers are afforded discretion as a necessary tool for responding to individuals' circumstances; they operate in domains where it is difficult to prescribe a clear course of action for every case. A central insight of Lipsky (2010) is that this discretion means that how policy is translated into practice is mediated through street-level bureaucrats, and that their collective behaviour constitutes a form of policy-making in its own right. How they exercise that discretion is shaped by the conditions of their work, above all ambiguous institutional goals and resource scarcity, which lead workers to develop routines, simplifications, and informal priorities as ways of managing demands that can distort the overarching goals pursued by an institution (Evans and Harris 2004; Lipsky 2010).

An immediate consequence is that street-level bureaucracies, by their very nature, are difficult to manage from above (Brodkin 2012). Institutional goals are often ambiguous, reflecting political compromise across competing dimensions rather than clearly speci-

[11] The history of formalization in public institutions continues through the development of expert systems (Simon 1997), the automation of welfare administration (Eubanks 2018), and more recent frameworks of evidence-based policymaking (Levy et al. 2021).

fied objectives (Coyle and Weller 2020), and the need to make decisions about individuals on a case-by-case basis makes it inherently difficult to develop performance measures that could effectively supervise how discretion is exercised. Such measures tend to capture what is easy to quantify, such as the number of cases processed, at the expense of dimensions that resist measurement, such as service quality; increasing throughput while service quality declines should not necessarily be seen as an improvement in institutional productivity (Lipsky 2010).

We may view algorithmic decision-making in public institutions as part of the history of managing and constraining street-level discretion (Alkhatib and Bernstein 2019; Vredenburgh 2025). Algorithms promise to reduce unwanted arbitrariness in decision-making, but they inherit similar difficulties. If the right decision is hard to specify as a performance measure for caseworkers, it may be equally hard to formalize as an objective for a predictive system. Lipsky already remarked that "the nature of service provision calls for human judgment that cannot be programmed and for which machines cannot substitute" (Lipsky 2010, p. 161) and Alkhatib and Bernstein (2019) argue that "street-level algorithms" may not be equipped to remain responsive to individual circumstances. If algorithmic tools are understood not as replacements for caseworkers but as decision aids, then their effects depend on how caseworkers actually use them. Street-level bureaucrats have a long history of responding to managerial interventions in unanticipated ways (Brodkin 2012), and there is little reason to expect algorithmic tools to be an exception.

The difficulties of integrating algorithms into public institutions can also be understood in the tradition of policy implementation studies, which have long examined how well-intentioned interventions fail along with the gap between policy design and execution (Pressman and Wildavsky 1984; Signé 2017; Hill and Hupe 2022). Lipsky (2010) can be read as part of a broader critique of top-down perspectives that impose often underspecified high-level goals while neglecting institutional context. Rittel and Webber (1973) famously call out the "wicked" nature of many social policy problems, which are ill-defined and open-ended, and involve multiple stakeholders with competing interests, resisting the kind of clean formalization that algorithmic approaches often presuppose (Alford and Head 2017). Liu et al. (2026) can be seen as informed by this tradition, scoping what an implementation science of modern automated decision-making might look like.

Much of this discussion has framed algorithmic decision-making as an intensification of earlier efforts to introduce mechanical rules into public institutions, subject to familiar tensions around goal ambiguity, professional discretion, and the political and social processes surrounding formalization. To close, it is worth asking what, if anything, may be qualitatively new about the application of predictive tools in these settings[12]. The answer will depend on the specific "bureaucratic counterfactual" one has in mind (Johnson

[12] We remain centered on predictive tools here. How large language models and related technologies may reshape this landscape is discussed in the outlook of this dissertation.

and Zhang 2022), ranging from categorical decision-rules to relatively unconstrained case-worker discretion. However, we can make a few general observations.

First, modern ML models are generally understood to be more opaque than their technical antecedents (Engstrom et al. 2020), creating a tension with the history traced above, in which quantification was often championed precisely as a tool for making government more accountable (Porter 1995). What kind of transparency these tools provide, and what kind we should demand, remains an open question (Amarasinghe et al. 2023). At the same time, algorithmic systems may require a more explicit formalization of objectives than previous systems, providing a "forced" transparency about institutional goals (Coyle and Weller 2020). The scale and proliferation of such tools in government is also reaching new heights (Levy et al. 2021), with rising capabilities allowing them to be embedded in an expanding range of functions (Engstrom et al. 2020). This may shift power structures, by replacing higher-level deliberative decision-making in agencies while also encroaching on the territory of classical street-level bureaucrats (Engstrom et al. 2020; Johnson and Zhang 2022). ML models can represent more complex and nonlinear decision rules than manually deliberated systems (what Johnson and Zhang (2022) call the "multidimensionality of need"), allowing for more personalization in their responses to individual cases (Levy et al. 2021). While these tools may be individually responsive in ways that Lipsky (2010) did not anticipate, they also require a more explicit formalization of goals that may collapse the ambiguity and exceptions previous systems were able to accommodate (Johnson and Zhang 2022; Kiviat et al. 2026).

2.5 Allocation of Societal Resources

As we have seen, prediction models in the public sector are often used to guide the allocation of scarce resources. These are settings in which goods such as healthcare, housing, welfare support, and employment services are not allocated through markets, often because there is a broader societal consensus that access should not depend primarily on willingness or ability to pay (Tobin 1970; Roth 2007). Instead, institutions must make explicit design choices about how such resources are distributed and which criteria should govern access. Questions of this kind have a long tradition across economics, computer science, and political philosophy (Elster 1992; Freedman et al. 2020; Das 2022).

Broadly, an allocation problem can be characterized by the population served, the resources allocated, the constraints imposed, and the objective the institution is pursuing.

- *Population*. In the settings we consider, resources are typically allocated to individuals (or cases (Kiviat 2023)) interacting with public institutions, such as job seekers registering at an employment office or households receiving a cash transfer. But one could consider other levels of aggregation, such as distributing budgets across schools or employment offices.

- *Resources.* Allocated resources can be homogeneous or heterogeneous, and divisible or indivisible, ranging from interchangeable ICU beds and uniform program slots to cash transfers of variable size or a choice among different support programs (Das 2022; Kuppler et al. 2022). What is allocated may represent a benefit or a burden (Elster 1992) to the individual, the difference between what Brayne (2017) and Johnson and Zhang (2022) call "algorithms of care" and "algorithms of suspicion".

- *Constraints.* Resources are typically scarce, meaning they cannot be provided to everyone. The exact constraint structure depends on the specifics of the problem, including whether one or multiple types of resources are allocated, whether decisions are made in a single batch or dynamically as individuals arrive over time, and whether every individual must receive some resource or only a subset is served.

- *Objectives.* How institutions choose to allocate resources is often (at least implicitly) underpinned by a principle of distributive justice (Kuppler et al. 2022). Elster (1992) provides a descriptive account of different principles of *local justice*[13] that inform how different non-market institutions distribute resources. Elster (1992) broadly distinguishes between principles that allocate on the basis of some property of the individual and those that do not. The latter, such as lotteries or queuing, are not conditioned on individual characteristics; the former rank recipients by individualized *criteria* such as need, desert, or expected benefit (Das 2022; Kuppler et al. 2022). Principles can also overlap (Elster 1992); for instance, Jain et al. (2024) argue for lotteries weighted by a predicted criterion.

Here, we are mostly concerned with simple allocation problems, in which a scarce, homogeneous good is allocated to a batch of individuals according to individual decision criteria. More specifically, we consider allocation principles that imply a utility function u that encodes how the planner values allocating the resource to an individual given their decision-relevant characteristics (Kuppler et al. 2022). In the simplest case, where u depends on a single observed outcome y, the allocation problem takes the following form:

Definition 2.5.1 (Allocation Problem)**.** Let $\mathcal{D}$ be a distribution over pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ and let $\alpha \in (0, 1)$ denote a capacity constraint. Fix a utility function $u : \mathcal{Y} \times \{0, 1\} \to \mathbb{R}$. The social planner's problem is then to find a policy $\pi : \mathcal{X} \to \{0, 1\}$ that solves

$$\max_{\pi:\mathcal{X}\to\{0,1\}} \mathbb{E}[u(Y, \pi(X))] \quad \text{s.t. } \mathbb{E}[\pi(X)] \leq \alpha$$

[13] The term *local* reflects that these principles are set by relatively autonomous institutions, each within its own problem domain and national context. Justice in this sense is compartmentalized rather than *global* (Elster 1992).

This captures many triaging problems in the public sector (Kuppler et al. 2022), but leaves aside important related problems such as dynamic allocation, as in organ matching (Das 2022) or housing allocation (Azizi et al. 2018); matching problems with richer constraints and incentive structures, such as, student assignment to schools (Abdulkadiroğlu and Sönmez 2003); and fair division problems with individual preferences (Amanatidis et al. 2023).

2.5.1 Solving Allocation Problems

As discussed in Section 2.3, allocation problems are often solved in two stages: one first estimates a model of the relevant outcomes (e.g., $f(x) \approx \mathbb{E}[Y \mid X = x]$), then plugs these estimates into the expected utility and solves the resulting optimization problem (Kube et al. 2023; Baldeschi et al. 2025; Liu et al. 2026). We will make use of simple plug-in approaches throughout this dissertation, as our focus lies less in improving methods for solving a given allocation problem than in the assumptions underlying its design.

However, treating the predictions as if they were correct does not necessarily lead to a good allocation (Fernandez-Loria and Provost 2022b). Models are typically trained to minimize a loss that differs from the downstream welfare, and since predictions are rarely perfect, we may have few guarantees on the expected utility of the resulting allocation. A predictor that looks good by its own loss can still induce a poor allocation if its errors fall in the parts of the population that determine the decision. A broad literature studies the link between prediction and decision-making, and how allocation problems might be solved more directly; we briefly survey some relevant strands here.

Calibration Calibration has long been associated with a predictors being "trustworthy" for downstream decision-making (Foster and Vohra 1997; Zhao et al. 2021). It requires that, conditional on any given forecast, the average realized outcome agrees with that forecast. Concretely, for any value v in the range of f,

$$\mathbb{E}[Y \mid f(X) = v] = v.$$

When the utility is linear in the outcome, a perfectly calibrated predictor lets a decision-maker trust the expected utilities implied by any decision rule that operates on its forecasts; in this sense, taking predictions at face value is optimal for maximizing downstream utility among the policies that map forecasts to actions (Zhao et al. 2021; Kiyani et al. 2025). A calibrated predictor can nonetheless be uninformative, in which case the resulting allocation may remain poor. For example, a constant predictor that simply returns the average population outcome is calibrated, but not particularly helpful for a decision-maker. To be competitive with richer policy classes that map covariates directly to actions, we need stronger guarantees on the predictions, such as multicalibration (Hebert-Johnson et al. 2018).

Because achieving full calibration is often difficult (high-dimensional outcome spaces are a case in point), recent work studies weaker notions that require calibration only with respect to the decisions being made, and shows that these can still recover similar downstream guarantees (Zhao et al. 2021; Kiyani et al. 2025). For constrained settings, Hu et al. (2023) extend omnipredictors to a range of downstream constrained loss-minimization tasks, while Baldeschi et al. (2025) study multicalibration guarantees for predictors used in weighted-graph matching.

Decision-Focused Learning We may also want to train a predictor directly on the decision loss (Liu et al. 2026). This matters most when the optimization has a structure the learner can exploit, like a complex set of constraints, and a utility under which some prediction errors are far more costly than others. Training this way is difficult because it is typically hard to differentiate through the downstream optimizer (Wilder et al. 2019). Different strategies have been proposed to address this difficulty; Elmachtoub and Grigas (2022), for example, suggest constructing a surrogate decision loss. These works as well as Kotary et al. (2021) and Mandi et al. (2024) offer a broader overview of the field. Parts of the learning-to-rank literature connect to this line of work, with ranking problems being particularly relevant for triage. For example, pairwise losses penalize misordered pairs, while listwise losses consider the whole ranking at once (Li 2015). Zehlike et al. (2022a) and Zehlike et al. (2022b) offer useful insights into how fairness constraints can be integrated for ranking problems.

Policy Learning Learning individualized treatment rules has a long history across econometrics, statistics, and computer science (Manski 2004; Kitagawa and Tetenov 2018; Athey and Wager 2021; Wager 2025). The goal is typically to find a policy that maximizes the expected welfare $V(\pi) = \mathbb{E}[Y(\pi(X))]$.[14] We can again take a plug-in approach (Qian and Murphy 2011; Wager 2025), first estimating the conditional treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ and then intervening whenever the estimated effect is positive.

$$\pi(x) = \mathbf{1}\{\hat{\tau}(x) > 0\}$$

Most directly relevant to Definition 2.5.1, Bhattacharya and Dupas (2012) estimate the CATE non-parametrically, threshold it to respect a budget constraint, and derive asymptotic welfare guarantees for the resulting rule.

In practice, however, we often want the policy to come from a restricted class Π, like decision rules with a simple functional form (Kitagawa and Tetenov 2018). A common strategy is then to optimize an estimate of the policy value directly over the class,

$$\hat{\pi} \in \operatorname*{argmax}_{\pi \in \Pi} \hat{V}(\pi)$$

[14] See Ben-Michael et al. (2024) for a discussion on policy learning with counterfactual utilities.

often called empirical welfare maximization (Kitagawa and Tetenov 2018; Athey and Wager 2021; Wager 2025). This estimate can be formed in different ways. For example, inverse-propensity weighting or doubly robust scores can be used (Kitagawa and Tetenov 2018; Bennett and Kallus 2020; Athey and Wager 2021).

For many commonly used policy classes, this is a non-convex optimization problem, since the objective depends on the policy parameters only through binary treatment assignments. In many cases, the optimization problem is solved by recasting it as a weighted classification problem (Zadrozny 2003; Zhao et al. 2012). The regret guarantees of empirical welfare maximization, however, are derived for the exact problem and do not necessarily carry over to common practical relaxations, in particular when the 0-1 classification objective is replaced by a convex surrogate loss (Bennett and Kallus 2020; Kitagawa et al. 2021; Wager 2025).

Wager (2025) provides a solid introduction to policy learning from observational data, an active research area that also covers additional desiderata, such as fairness constraints (Frauen et al. 2024) and ways to deal with unobserved confounding (Kallus and Zhou 2018).

2.5.2 Rethinking the Role of Prediction in Allocation

Arguably, a more accurate prediction model will allow an institution to target its resources more precisely, ideally improving the welfare of the people it serves. But better prediction is only one of many policy interventions that could do so (Liu et al. 2026). For example, the institution might instead improve the quality of the resources it allocates, or simply serve a larger share of the population. Building better predictive systems also tends to be costly, requiring additional data collection, technical infrastructure, and scarce institutional attention, so the value of better prediction must be weighed against the other ways those resources could be used. Liu et al. (2023), for instance, cite education experts who note that even accurate predictions of dropout risk do little good if the school lacks the resources to actually help all the students identified.

Taking up this question, a growing body of work contextualizes the value of prediction and individual targeting in social allocation problems, drawing on both domain-specific empirical analysis and theoretical work with stylized models.

For example, we may question the value of individual-level prediction in the first place, and ask when it is worth sticking with the "bureaucratic counterfactual" (Johnson and Zhang 2022) instead. For instance, a simpler decision rule can be easier to interpret and cheaper to run, requiring less data infrastructure and modeling work. Perdomo et al. (2025) study early warning systems in schools and find that a simple rule based on the academic environment performs as well as predicting individual dropout risk; the more pressing problem, they suggest, is finding better educational interventions. On the theoretical side, Shirali et al. (2024) formalize this intuition, arguing within a stylized model that individual-level prediction pays off only when inequality between aggregate units is low and the budget is large. Such conclusions can be brittle, as they are sensitive to the un-

derlying assumptions and vary across domains. Mashiat et al. (2025) find that individual prediction substantially outperforms neighborhood targeting for door-to-door outreach to prevent tenant eviction, in part because the costs have a spatial structure. Similarly, Aiken et al. (2022) find that ML-based targeting outperforms geographic targeting of humanitarian aid in Togo.

Other work directly formalizes the tradeoff between improving prediction and investing in the other components of the system. For example, Aiken et al. (2025) compare the cost-effectiveness of different targeting strategies for cash transfer programs, and Hanna and Olken (2018) weigh universal against targeted transfer programs. In a related setting, Aiken et al. (2023) study how frequently a poverty prediction model should be retrained, given that data collection is costly and competes with the transfer budget.

Most closely related to this dissertation, Perdomo (2024) analyzes the *relative value of prediction* when allocating social goods, comparing the marginal welfare impact of slightly improving a simple predictor to that of expanding the number of individuals served. He formalizes this tradeoff by introducing the *prediction-access ratio* under a simple linear model. Part of this dissertation (Chapter 6) builds on this line of work, investigating similar tradeoffs when targeting the worst-off (i.e., those most in need) and conducting an empirical analysis with governmental data from the German employment service (Fischer-Abaigar et al. 2025).

Taken together, and perhaps unsurprisingly, all of this work shows that the value of prediction is context-dependent, shifting with cost structures, utilities, and the available competing investments. Building on this observation, we situate improvements in prediction within a broader meta-design space of allocation problems (Fischer-Abaigar et al. 2026), alongside the other parameters that define an allocation, such as the capacity and utility of the resource allocated, and provide tools to prioritize investments across them (Chapter 7).

Most of the work surveyed in this section takes the underlying (welfarist) allocation principle as given, and asks how to pursue it more efficiently, usually from the perspective of a central social planner. One could go further and question the principle itself, asking when allocation based on individualized criteria is a good idea at all, and whether other mechanisms, such as lotteries, might be preferable and more justified (Stone 2007; Jain et al. 2024).

3 Overview of Contributions

This section introduces the four contributing papers of this dissertation. Taken together, they examine the role of predictive systems within the broader institutional settings in which they are embedded and ask two related questions. First they ask whether the assumptions underlying a predictive system match the deployment context it operates in, a prerequisite for determining whether the system is actually doing what it claims. Second they investigate whether improving prediction is the most valuable investment in the first place, given the range of alternative actions a social planner might pursue to raise downstream welfare.

In the first paper of this dissertation (Chapter 4), we examine how predictive systems can become misaligned with the public sector institutions in which they are embedded. We argue that many failures of such systems stem from implicit assumptions made during technical implementation that do not hold in dynamic institutional environments. On the input side, the assumptions underlying the data may fail through distribution shifts, label bias, and the influence of past decision-making on the available outcomes. On the output side, predictive systems often impose a narrow framing on the problem, making it difficult to capture the multiple and potentially competing policy objectives at stake, or to integrate the human decision-makers who act on their output. In response, we argue that modeling efforts should center more overtly on improving decision-making outcomes rather than on predictive accuracy alone, which will also require shifting from purely predictive models towards other (causal) estimands and modeling frameworks. We discuss the assumptions underlying these different modeling approaches and survey methodological work that could help address some of the challenges raised, for example dealing with distribution shifts. We close by discussing where external input from domain experts and stakeholders is needed to guide these choices.

While much has been written on the tensions between AI and public sector decision-making, we found it useful to show how simply making the assumptions behind many predictive systems explicit already reveals a disconnect with the environment in which they operate, one that makes it hard to anticipate how (and if) a system will perform well. In this sense the paper complements the literature in Section 2.4.2, such as Coston et al. (2023) and Wang et al. (2023), which share a similar framing and ask what the necessary conditions for safe deployment and evaluation might be. It also connects to later work by Liu et al. (2026), who discuss predictive systems as policy interventions to be evaluated within their institutions.

The second paper (Chapter 5) is a shorter, didactic comment, arguing that causal reasoning is a necessary, though not sufficient, condition for reliable algorithmic decision-making. It picks up and sharpens an argument from the first paper, namely that even purely predictive decision rules carry implicit causal assumptions, so the question is not whether one makes causal assumptions but whether one does so explicitly. It also examines how open challenges to reliability, such as data quality, uncertainty quantification, and performativity, interact with a causal perspective, and points to directions for future research.

The first strand points to a lesson that seems obvious but is easily forgotten: better predictions do not necessarily translate into better institutional decisions. If a model is built on assumptions that do not hold when it is deployed, it becomes difficult to anticipate how it will perform or to evaluate whether it does what it claims. The second strand takes a different angle on the role of prediction in its wider operational context, namely whether it is worth improving in the first place. Once we accept that a predictive system is embedded in a broader institutional setting, that setting also implies a design space of alternatives a planner might pursue to raise downstream welfare, such as expanding the capacity of a program or improving service quality.

In the third paper (Chapter 6), we investigate the relative value of prediction in a well-specified allocation problem in which an institution uses a simple risk predictor to prioritize the worst-off in a population. Specifically, we ask how a marginal improvement in prediction accuracy compares to a marginal expansion of bureaucratic capacity, that is, serving a slightly larger fraction of the population. We approach this through a stylized Gaussian model, and find that in resource-constrained settings the relative impact of expanding capacity typically far outweighs the welfare impact of improving prediction. Prediction is particularly valuable as a first- and last-mile effort: when the baseline system explains either none or almost all of the variance in outcomes. The mathematical results build on the notion of the *prediction-access ratio* introduced by Perdomo (2024), which we extend to the setting of targeting the worst-off (see Section 2.5.2). We complement the theoretical analysis with a case study using administrative data on hundreds of thousands of job seekers from the German employment agency, finding qualitatively similar results.

However, the tradeoff between prediction and capacity is only one of many design choices that shape how well an allocation system serves its population, and the external validity of theoretical results derived in stylized settings necessarily depends on how closely those settings match the problem at hand. In the fourth paper (Chapter 7), we formalize and motivate the broader meta-design space of allocation problems. Many parameters of an allocation that are typically treated as fixed are ultimately design choices made by a social planner, and matter as much as, if not more than, how best to solve a given problem. We show how a planner can navigate this space by empirically evaluating the *welfare surface*, which describes how the optimal value of a targeting rule varies across design configurations, across different capacity constraints, data distributions, and utilities. Many of the key tradeoffs can be read off the geometry of this surface, including which configu-

rations are optimal under a given budget and which design axes matter most at a current configuration. Rather than providing universal answers, the framework gives planners tools to derive conclusions that will be valid in their own context. We demonstrate this on two real-world datasets: the targeting of cash transfers in Ethiopia and the profiling of job seekers in Germany. We release a small software toolkit RVP[1] to support this type of analysis.

[1] https://github.com/unai-fa/relative-value-of-prediction

Part II

Publications

4 Bridging the Gap

Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

Publication

U. Fischer-Abaigar, C. Kern, N. Barda, and F. Kreuter (2024). "Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector". *Government Information Quarterly* 41:4, p. 101976. ISSN: 0740624X. DOI: 10.1016/j.giq.2024.101976

Author Contributions UFA led the project as first author, conceptualized the paper, designed its structure and figures, and wrote the original draft. CK contributed to conceptualization and writing and supervised the project. NB and FK contributed to reviewing and editing, and FK co-supervised the project.

Government Information Quarterly 41 (2024) 101976

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector

Unai Fischer-Abaigar [a,b,*], Christoph Kern [a,b,c], Noam Barda [d,e], Frauke Kreuter [a,b,c]

[a] Dept. of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany
[b] Munich Center for Machine Learning, LMU Munich, Geschwister-Scholl-Platz 1, Munich, Germany
[c] Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, 7251 Preinkert Dr., College Park, MD 20742, USA
[d] Dept. of Software and Information Systems Engineering, Ben Gurion University of the Negev, Be'er Sheva, Israel, 1 Ben-Gurion Ave, 8410501 Be'er Sheva, Israel
[e] Department of Epidemiology, Biostatistics and Community Health Sciences, Ben Gurion University of the Negev, 1 Ben-Gurion Ave, 8410501 Be'er Sheva, Israel

ARTICLE INFO

Keywords:
Automated decision-making
Reliable artificial intelligence
Public policy
Causal machine learning

ABSTRACT

AI-driven decision-making systems are becoming instrumental in the public sector, with applications spanning areas like criminal justice, social welfare, financial fraud detection, and public health. While these systems offer great potential benefits to institutional decision-making processes, such as improved efficiency and reliability, these systems face the challenge of aligning machine learning (ML) models with the complex realities of public sector decision-making. In this paper, we examine five key challenges where misalignment can occur, including distribution shifts, label bias, the influence of past decision-making on the data side, as well as competing objectives and human-in-the-loop on the model output side. Our findings suggest that standard ML methods often rely on assumptions that do not fully account for these complexities, potentially leading to unreliable and harmful predictions. To address this, we propose a shift in modeling efforts from focusing solely on predictive accuracy to improving decision-making outcomes. We offer guidance for selecting appropriate modeling frameworks, including counterfactual prediction and policy learning, by considering how the model estimand connects to the decision-maker's utility. Additionally, we outline technical methods that address specific challenges within each modeling approach. Finally, we argue for the importance of external input from domain experts and stakeholders to ensure that model assumptions and design choices align with real-world policy objectives, taking a step towards harmonizing AI and public sector objectives.

1. Introduction

Automated decision-making (ADM) systems are increasingly being adopted across the public sector (Chiusi, 2020; Mitchell, Potash, Barocas, D'Amour, & Lum, 2021; Levy, Chasalow, & Riley, 2021), often relying on AI models to address a wide array of problem domains, including critical areas such as predictive policing (Lum & Isaac, 2016), criminal justice (Angwin, Larson, Mattu, & Kirchner, 2016; McKay, 2020), fraud detection in government (Engstrom, Ho, Sharkey, & Cuéllar, 2020), child abuse prevention (Chouldechova, Benavides-Prado, Fialko, & Vaithianathan, 2018), tax audit selection (Black, Elzayn, Chouldechova, Goldin, & Ho, 2022), early warning systems in public schools (Perdomo, Britton, Hardt, & Abebe, 2023), credit scoring (Kozodoi, Jacob, & Lessmann, 2022), profiling of job seekers (Bach, Kern, Mautner, & Kreuter, 2023; Desiere & Struyven, 2021; Körtner & Bonoli, 2023), development aid (Kuzmanovic, Frauen, Hatt, & Feuerriegel, 2024) and public health (Potash et al., 2015). Despite expectations of enhancing decision-making by improving reliability, objectivity, efficiency and uncovering factors that traditional institutional processes may overlook, ADM systems face considerable challenges (Barocas, Hardt, & Narayanan, 2023; Coston, Kawakami, Zhu, Holstein, & Heidari, 2023; Engstrom et al., 2020; Levy et al., 2021; Wang, Kapoor, Barocas, & Narayanan, 2023). Real-world examples demonstrate shortcomings, ranging from racial and gender bias to systems exhibiting poor predictive accuracy leading to flawed decision-making (Allhutter, Cech, Fischer, Grill, & Mager, 2020; Angwin, Larson, Mattu, & Kirchner, 2016; Dressel & Farid, 2018; Mayer, Strich, & Fiedler, 2020; Obermeyer, Powers, Vogeli, & Mullainathan, 2019). Such unintended consequences

* Corresponding author at: Dept. of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany.
E-mail addresses: unai.fischerabaigar@stat.uni-muenchen.de (U. Fischer-Abaigar), christoph.kern@stat.uni-muenchen.de (C. Kern), noambard@bgu.ac.il (N. Barda), frauke.kreuter@stat.uni-muenchen.de (F. Kreuter).

https://doi.org/10.1016/j.giq.2024.101976
Received 22 April 2024; Received in revised form 27 September 2024; Accepted 28 September 2024
Available online 11 October 2024
0740-624X/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

are particularly concerning due to their significant impact on individuals' lives and the potential reinforcement of systemic biases. Recent legislation, such as the European Union's AI Act, highlights these concerns by establishing regulations for high-risk AI systems (Laux, Wachter, & Mittelstadt, 2023).

A growing body of literature explores the challenges and potential benefits of employing AI systems to enhance decision-making within the public sector (Pencheva, Esteve, & Mikhaylov, 2020; Sun & Medaglia, 2019; Wirtz, Weyerer, & Geyer, 2019; Zuiderwijk, Chen, & Salem, 2021). Moreover, several reviews examine the adoption of AI in government (Levy et al., 2021), including US federal institutions (Engstrom et al., 2020) and the EU public sector (van Noordt & Misuraca, 2022; Matzat, 2019). These reviews cover a wide range of challenges, primarily focusing on institutional, ethical and legal implications of using ADM in the public sector.

In this article, we focus on challenges that arise from a misalignment between the technical assumptions underlying machine learning (ML) models and the realities of decision-making in complex public sector environments. Specifically, we will discuss AI-driven decision-making used for the allocation of scarce resources in the public sector, where decisions involve determining whether individuals qualify to receive specific interventions or services (Kuppler, Kern, Bach, & Kreuter, 2022). Our focus is on ADM systems that do not rely on manually encoded rules, but rather use supervised ML models to learn patterns from historical data to predict relevant outcomes that inform decision-making. Although ML approaches can vary widely, ranging from support vector machines to neural networks, we aim to keep our discussion relevant across different models by exploring the general limitations and challenges of using supervised ML for public sector decision-making. Throughout the text, we use terms like AI, ML and predictive algorithm interchangeably to refer to the computational model underlying the ADM system.

Decision-making in these environments often takes place in dynamic, evolving social contexts, which can conflict with the explicit formalization requirements demanded by ML models (Amarasinghe, Rodolfa, Lamba, & Ghani, 2023; Levy et al., 2021; Mitchell et al., 2021; Passi & Barocas, 2019). Technical choices made during model development often rest on implicit assumptions, such as stable data distributions and a straightforward link between prediction and decision-making, that may not hold in these complex settings. For example, policy objectives are often shaped by multiple stakeholders, political compromises and competing goals (Coyle & Weller, 2020; Levy et al., 2021), making it difficult to translate them into clearly defined objectives for ML systems. When the assumptions behind the technical model construction do not align with the deployment context, there is a risk of developing systems that fail to capture the complexities of real-world decision-making, potentially leading to adverse outcomes upon model deployment.

Consider, for example, a public employment service (PES) office that aims to determine which job seekers should participate in job programs to increase their re-integration chances. The PES wants to deploy ML to learn the optimal assignment of support to job seekers based on data collected as part of their daily operations. However, the PES now faces two critical sets of interconnected complications: first, while their data may include detailed records of labor market histories, they are operating in a complex and dynamic social environment which raises questions of distribution shift, feedback loops and the challenge of accounting for the effect of competing (current) and previous job support programs. Second, the PES needs to ensure that the predictions can effectively be integrated in their current decision-making practices. This may require model guarantees to build caseworker trust in the predictions and ensuring that other relevant objectives and constraints are sufficiently incorporated in the system. All these issues require careful consideration in the model design choices. A misalignment between technical assumptions and problem setting, such as building a model under the implicit assumption that labor market characteristics remain invariant, may result in unintended consequences, such as an allocation mechanism that might become unreliable over time.

Efforts to analyze challenges from a technical perspective are ongoing and focus on connecting methodological AI research with the unique demands of high-stakes decision-making. These efforts explore various subdimensions of this complex issue, including training data quality (Shahbazi, Lin, Asudeh, & Jagadish, 2023), target variable bias (Guerdan, Coston, Wu, & Holstein, 2023) and uncertainty (Gruber, Schenk, Schierholz, Kreuter, & Kauermann, 2023; Kaiser, Kern, & Rügamer, 2022). Furthermore, active research develops frameworks to examine the conditions under which the usage of predictive algorithms for high-stakes decision-making can be justified (Coston et al., 2023; Wang et al., 2023).

In this work, we identify and analyze misalignments that commonly occur between ML models and public sector decision-making. Guided by the 'ADM process model' (Gerdon, Bach, Kern, & Kreuter, 2022), we focus on how models connect with their wider real-world deployment context by examining both the data assumptions (model input) and how models are integrated into the decision-making process (model output). Using the lens of misalignment developed here, we build on the recent technical literature on ML and decision-making to isolate five specific challenges that we consider to exemplify the type of issues that can occur at these two interfaces: distribution shift, label bias and the influence of past decision-making on the input side, and competing objectives and constraints and human-in-the-loop interactions on the output side. We analyze each of these challenges to better understand how misaligned technical assumptions can lead to erroneous decision-making and adverse outcomes for affected individuals in public sector environments.

Through our analysis, we find that standard ML methods often rely on assumptions that do not fully account for the complexities of public sector decision-making. In response, we propose a shift in modeling efforts from focusing solely on predictive accuracy to improving decision-making outcomes. We argue that achieving this shift may, in certain cases, require alternative modeling techniques that extend purely predictive models, and more directly center on the goal of decision-making. With this in mind, we highlight promising developments in causal machine learning, including counterfactual prediction and policy learning. Within each modeling framework, we summarize technical methods that provide (partial) solutions to the identified challenges. To guide practitioners in selecting the right approach, we clarify the assumptions underlying each framework, specifically addressing, how the model estimand connects to the utility of the decision-maker and the data and assumptions required for reliable estimation.

By examining these frameworks through the lens of public sector decision-making, we want to encourage technical practitioners to carefully consider the assumptions behind different modeling approaches and expand their toolbox to include methods that may be better suited for complex, real-world decision-making. For policymakers, domain experts and other stakeholders, we outline which external input is important to help model developers make the right assumptions to inform model design.

While we do discuss several risks resulting from the assumptions made during model development, zooming out to the institutional and societal context raises more complex issues. For instance, institutional and cultural biases embedded in historical data as well as data collection methods and processing can significantly contribute to discriminatory decision-making (Fountain, 2022; Janssen & Kuk, 2016). Algorithmic systems may also reinforce existing structural inequalities by formalizing problematic decision-making practices (Kolkman, 2020) or empowering institutions with unjust goals. The continued digitization of bureaucratic processes, particularly when multiple institutions and systems interact, can create new risks, such as making it harder to correct errors across systems or systematically excluding specific user groups (Peeters & Widlak, 2018). However, we consider addressing misalignments between assumptions made during model development

4 Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

and the deployment context to be essential for avoiding harmful model design, making it a necessary (though not sufficient) condition for the fair development of AI-driven decision making in the public sector.

This paper is structured as follows. We first explore central (mis) alignment challenges that occur along the ML pipeline when developing and deploying AI systems to support decision-making in the public sector (Section 2). Second, we highlight recent methodological developments that exceed the classical supervised ML paradigm, showing promise in addressing the challenges identified (Section 3). Third, we discuss the selection of an appropriate modeling approach in a given deployment context (Section 4). In the discussion, we address broader issues related to ADM in the public sector, specifically highlighting the importance of domain expertise and stakeholder input (Section 5). Finally, Section 6 provides a concise summary of our findings.

2. Defining the gap: Central challenges in connecting ML and decision-making

Predictive models for ADM systems are designed to inform decisions in (interaction with) dynamic social contexts, which gives rise to a list of fundamental challenges. This includes questions related to choosing adequate model input, as the effectiveness of any ML model is fundamentally linked to the quality of its training data. Ensuring that this data accurately represents the target population is key to avoid biased and unreliable predictions (Gruber et al., 2023). Securing representative data in the public sector, however, is a complex task. Public sector data is incredibly diverse (Dwivedi et al., 2021; Janssen, van der Voort, & Wahyudi, 2017), comes in a variety of formats, often lacks structure and encompasses a wide array of data modalities (Dwivedi et al., 2021). Despite the abundance of data in theory, high-quality data suitable for ML is often not easily available in the public sector (Alexopoulos et al., 2019; Sun & Medaglia, 2019). In many countries, the lack of robust infrastructure to enable data sharing and integration of various data sources can hinder the development of ML models (Sun & Medaglia, 2019; Wirtz et al., 2019), stemming from issues such as resource constraints, data protection, safety concerns and institutional pushback. These issues are especially concerning for ADM systems, since building a model that is capable of informing future decision-making places significant demands on the training data (Coston et al., 2023; Hüllermeier, 2021).

In addition, it is important to consider how the model output will be integrated into the decision-making process. Rather than improving model performance in isolation, the success of a system should be evaluated based on whether it helps guide decision-making to achieve the intended policy objectives (Mitchell et al., 2021). Choosing the appropriate modeling setup, requires drawing a connection from broad, often hard to formalize policy goals to the specific target outcomes estimated by the prediction model (Levy et al., 2021).

ADM systems aim to identify individuals for targeted interventions, typically with the goal of improving an objective defined as the aggregate of the individual outcomes of interest. For instance, policy makers may seek to improve healthcare in a hospital as a function of individual treatment outcomes or maximize money recovered during tax audits (Black et al., 2022). These overarching policy goals may be formalized through an allocation principle that determines the optimal assignment of interventions based on the estimated outcomes (Kuppler et al., 2022). For example, we may choose to intervene when the effect of an intervention is expected to be positive (Fernández-Loría & Provost, 2022a), or apply an intervention only for the k top-ranked individuals based on their (predicted) individual outcomes of interest (Amarasinghe et al., 2023; Kuppler et al., 2022). The last approach reflects real-world resource constraints typical in the public sector, for example a limited number of staff and financial resources. However, the link between intended goals of a system, allocation principle and prediction targets is often more complicated than this setup suggests. Often additional goals and information needs to be considered before a final decision is made, such as the opinion of a human decision-maker.

In the following subsections, we discuss key challenges associated with both data input and model output that are especially relevant for high-stakes decision-making in the public sector (see Fig. 1). These challenges include considering potential distribution shifts between the model's training and deployment context (Section 2.1), dealing with proxy variables in complex policy settings (Section 2.2) and discerning the impact of past decision-making on the data (Section 2.3). We will also discuss the difficulty of handling multiple potentially conflicting goals (Section 2.4), and the role of human decision-makers whose judgment can potentially overrule the recommendations made by an algorithm (Section 2.5).

2.1. Distribution shifts

A key challenge when using ML models is to ensure that they perform well in real-world situations. Often, the data used to train and evaluate the model will not fully represent the actual population in the environment where the model will be deployed (Gruber et al., 2023; Kouw & Loog, 2018). This mismatch between the distribution of training and deployment data is commonly referred to as distribution shift, and can lead to a significant overestimation of a model's performance (Gruber et al., 2023; Kouw & Loog, 2018). In other words, the model may learn patterns in the training data that do not generalize well to the deployment data, causing it to perform poorly in practice. Distribution shifts are especially challenging in the public sector, where sourcing reliable data can be particularly difficult. Models are often deployed in complex and evolving social contexts, and limited resources, such as a shortage of technical staff (Wirtz et al., 2019), make it difficult for public institutions to monitor model performance for unexpected distribution shifts.

There are several types of shifts that can occur (Kouw & Loog, 2018; Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, & Herrera, 2012). For example, the distribution of input covariates, such as age, income or educational background, might vary if a model is trained in one geographic region but deployed in another. An even harder type of shift to address is when the relationship between input covariates and the outcome changes, making the model's predictions less applicable in the new setting.

Distribution shifts often result from a biased selection of training data (Moreno-Torres et al., 2012). For example, it may be more costly to collect relevant data for hard to reach subgroups in the population (Tourangeau, 2019), leading to them being underrepresented in the data used for training. Selection bias may be introduced through a variety of other mechanisms, for example if past decision policies have led to only certain subgroups receiving an intervention, it will be difficult to assess the interventions' impact for other individuals.

Even a comprehensive selection of training data does not guarantee long-term robustness. As changes in the deployment environment occur, the initially accurate data may become increasingly outdated, likely causing the performance of a model to degrade over time (Moreno-Torres et al., 2012). For instance, labor market characteristics might change over time, making a model trained on older data for predicting unemployment less accurate. When a model is used to inform future decision-making, its continued deployment may itself be a source of distribution shift. For instance, individuals might strategically manipulate attributes that are not causally related to the true outcome but are correlated to improve predictions in their favor, often worsening the model's accuracy in predicting the true outcome of interest (Hardt, Megiddo, Papadimitriou, & Wootters, 2016). This is a known challenge when making use of models to support enforcement decisions in government, such as financial fraud detection. In order to evade detection, certain regulatory subjects will adapt to a given system, thereby requiring continuous updates to maintain effectiveness (Engstrom et al., 2020). The performance of a model may also be impacted by more sudden changes in the deployment environment. The introduction of a

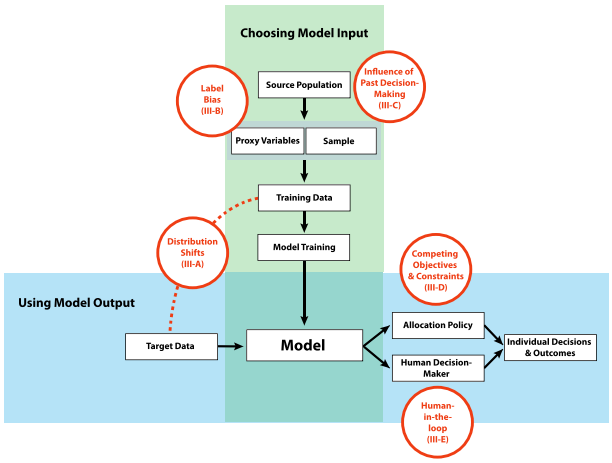


Fig. 1. Overview of the Primary Technical Challenges at the Intersection of Public Sector Decision-Making and Machine Learning. The challenges (highlighted in red) are positioned along the ML pipeline, with emphasis on data collection and model training (green) and model deployment to support decision-making (blue). For the sake of clarity, some overlapping challenges and connections have been omitted, such as the possible influence of decision outcomes on future data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

new policy can influence how new training data is collected and labeled, and unexpected events, such as the COVID-19 pandemic (Singh et al., 2021), can reduce the prediction accuracy of a model.

One common strategy to keep a ML model up-to-date is to regularly retrain it with new incoming data. However, caution is needed in scenarios where the model's output strongly influences future training data (Perdomo, Zrnic, Mendler-Dünner, & Hardt, 2020). Biased initial training data can result in self-fulling prophecies, in which model-informed interventions lead to new biased data that is fed back into the model training. Predictive policing is a canonical example of such a harmful feedback loop, where a higher police presence in neighborhoods classified as high-risk by the model can lead to higher arrest rates (i.e. the proxy variable) independent of the true crime rate (Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2018).

To effectively anticipate distribution shifts, the insight of domain experts and stakeholders will often be key. For example, prior knowledge of which causal relationships between predictors and the outcome of interest are expected to remain invariant (Kerrigan, Hullman, & Bertini, 2021) can facilitate the selection of a dedicated approach to increase robustness under shifts. In general, prediction quality should be continuously monitored to detect signs of worsening model performance. This task goes beyond the technical challenges we will discuss, and necessitates dedicated institutional resources and procedures, such as requiring periodic re-approval of deployed systems (Levy et al., 2021).

2.2. Label bias

Obtaining accurate ground truth data in real-world settings is rarely simple, as the true quantity of interest is often not directly measurable (Barocas et al., 2023; Coston et al., 2023; Guerdan, Coston, Wu, & Holstein, 2023). While challenges in measuring outcomes are not unique to the public sector, they are particularly pronounced in the public sector, where projects often address complex social phenomena that are difficult to quantify such as health, social welfare and education. In contrast, the private sector typically evaluates outcomes using more straightforward metrics like return on investment (Wirick, 2011). These difficulties often encourage the use of proxy variables that are more easily available, such as hospitalization records and arrest rates. Similarly, it may take a considerable amount of time before the outcome of interest can be observed; predicting the 10-year risk of cardiovascular events takes at least a decade. Such lag may require the use of short-term outcomes as proxies (Athey, Chetty, Imbens, & Kang, 2019).

Using proxy variables can introduce bias into a model. Proxy variables often capture institutional responses rather than the true underlying outcome of interest. For example, using ICU hospitalization as a proxy for COVID-19 severity is an imperfect measure, as ICU admission will depend on other factors like bed availability and other admission criteria. This can be especially problematic if the relationship between proxy and true target varies by protected attributes, such as race and gender (Guerdan, Coston, Wu, & Holstein, 2023; Passi & Barocas, 2019). Obermeyer et al. (2019) demonstrate that using expected healthcare cost as a proxy for health needs in predictive algorithms can lead to significantly underestimating the risk score of Black patients. This is because Black patients with similar health needs generate fewer medical expenditures compared to white patients. Similar examples can be found in various application contexts, such as judicial bail prediction (Fogliato, Chouldechova, & G'Sell, 2020) and lending algorithms (Mitchell et al., 2021).

Mitigating such biases cannot be achieved by collecting more data; it demands careful consideration of the relationship between the the true label Y and the measured proxy label $\tilde{Y}$ (Gruber et al., 2023; Guerdan, Coston, Wu, & Holstein, 2023). Validating the assumptions made about the measurement process may require an evaluation of the proxy variable using external data. For example, in the evaluation of the Allegheny Family Screening Tool, an algorithm designed to aid in child maltreatment hotline screening, researchers utilized data from a pediatric hospital in form of hospitalization records to assess the relationship between the model's risk scores and the occurrence of injury encounters as recorded in the hospital's dataset (Cheng & Chouldechova, 2022;

4 Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

Vaithianathan, Kulick, Putnam-Hornstein, & Benavides-Prado, 2019).

2.3. Past decision-making

When developing an ADM system, we often encounter scenarios in which the available training data has been influenced by past decision-making (Coston, Mishler, Kennedy, & Chouldechova, 2020). For example, a model predicting the risk of job seekers becoming long-term unemployed with the aim of allocating future support programs needs to account for how such programs were distributed in the past. Otherwise the model will likely underestimate the risk of unemployment for individuals that used to receive prioritized support after the decision-making process is altered through the deployment of the model (Lenert, Matheny, & Walsh, 2019). The predictions of such a model would lead to misleading recommendations, since its predictions are valid only under the assumption that the decision-making policies remain unchanged or that the interventions were largely ineffective in the past (Dickerman & Hernán, 2020).

In such scenarios, it may be required to explicitly model the effect of interventions by predicting counterfactual outcomes, such as the expected outcome of a medical treatment for a specific individual. However, the estimation of counterfactual outcomes is difficult, as it relies on untestable assumptions due to the limitation of only observing one intervention outcome per individual. Causal modeling requires data on past interventions, specifically which interventions each individual was targeted with, whereas missing treatment data is common in real-world scenarios (Kennedy, 2020; Kuzmanovic, Hatt, & Feuerriegel, 2023). A central challenge in causal modeling are confounding variables, resulting in the group of individuals subjected to a specific intervention exhibiting systematic differences in outcomes compared to the overall population (Fernández-Loría & Provost, 2022a). Students from a more privileged socioeconomic background may find it easier to enroll in a free tutoring program, but may also tend to perform better on tests due to stronger support networks. This leads to a risk of overestimating the effectiveness of the program for students from the general population.

The canonical way of dealing with such bias are randomized controlled trials (RCTs) (Caron, Baio, & Manolopoulou, 2022). However, in many high-stakes public sector settings it will be impossible to conduct a RCT due to resource constraints and ethical limitations (Caron et al., 2022). When the efficacy of the interventions is well-established, a randomized study may be hard to justify, as in the case of criminal justice (Lakkaraju, Kleinberg, Leskovec, Ludwig, & Mullainathan, 2017) or child abuse prevention (Vaithianathan, Benavides-Prado, Dalton, Chouldechova, & Putnam-Hornstein, 2021).

Alternatively, causal outcomes may be estimated from observational data. However, this requires the assumption that relevant confounding variables have been observed, allowing for the disentanglement of past intervention assignment and outcomes. However, some variables may remain elusive (Lakkaraju et al., 2017; Rambachan, Coston, & Kennedy, 2022), such as the impressions gained by decision-makers from in-person interactions. In situations in which it is difficult to guarantee no unmeasured confounding variables, there still may be ways to estimate the outcome of interest. For instance, Chen, Li, and Mao (2023) and Lakkaraju et al. (2017) utilize data from multiple human decision-makers who were randomly assigned to cases to enable estimation.

2.4. Competing objectives and constraints

Formalizing the intended policy objectives into a clearly defined allocation principle is difficult, especially when dealing with multiple stakeholder groups, each with their distinct and potentially competing goals and constraints (Levy et al., 2021; Mitchell et al., 2021; Passi & Barocas, 2019). For example, a welfare agency may seek cost-efficient solutions, while ensuring fair decision-making. Similarly, when the IRS decides whom to audit, various objectives come into play, such as maximizing revenue, deterrence, and compliance with institutional and monetary constraints (Black et al., 2022). Regardless of the specific context, resource constraints are common in the public sector (Amarasinghe et al., 2023). These constraints may result from limited financial resources or be influenced by institutional factors, such as a limited workforce, legal regulations or external political considerations.

Predictive systems, however, typically encourage a more limited scope by estimating only one relevant factor (Mitchell et al., 2021). This singular focus may introduce omitted-payoff bias, a situation where a model target captures only a subset of critical objectives and constraints, potentially reducing the real-world utility of the system (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018). For example, the IRS disproportionately audits low-income owners compared to their high-income counterparts, despite higher misreporting of tax liability among the latter group (Black et al., 2022). While auditing low-income individuals is more cost-efficient, it can exacerbate social inequalities. This problem of narrow focus becomes especially pronounced when human decision-makers have fewer opportunities to incorporate additional considerations into the decision-making process and rely on the model's predictions too heavily.

Therefore, effort must be made to translate multiple policy goals and constraints into explicitly defined objectives for the ADM system (Coyle & Weller, 2020; Mitchell et al., 2021). The exact choice of the prediction target often represents a policy choice because it can have profound downstream impacts that should not solely be the responsibility of ML developers (Levy et al., 2021; Passi & Barocas, 2019). For instance, in the IRS example, shifting the prediction target from the probability of misreporting to predicting misreported income leads to a significantly more equitable distribution of audits, even without explicitly enforcing fairness constraints (Black et al., 2022). Integrating multiple goals into a system typically requires making explicit tradeoffs between different objectives and constraints. A familiar example of competing objectives during model development is that accuracy has to be sacrificed to enforce fairness constraints (Black et al., 2022; Kozodoi et al., 2022) or enhance model interpretability (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). In public policy, one common approach to assess competing goals is performing a cost-benefit analyses, valuing different impacts and objectives in monetary terms (Boardman, Greenberg, Vining, & Weimer, 2018). A similar approach in model design might permit the combination of multiple objectives into a single loss function. However, assigning a monetary value to different potentially incommensurable impacts or goals is not always straightforward, resulting in critiques of this utilitarian approach to decision-making (Hwang, 2016).

Clearly specifying the optimization targets of an ADM system is a delicate process that carries the risk of distorting the originally intended goals (Levy et al., 2021). This risk is especially pronounced when some policy goals are easier to formalize than others, prompting the oversimplification of complex issues through an algorithmic lens (Levy et al., 2021). Stakeholders' preference for cost-effective, straightforward solutions and easily measurable prediction targets may exacerbate this problem (Barocas et al., 2023). Nevertheless, decisions must be made, and the growing use of ML in the public sector will likely require new dialogues among stakeholders, while also providing an opportunity to make the weighting and tradeoffs between policy objectives more explicit and transparent than in the past (Coyle & Weller, 2020; Levy et al., 2021).

2.5. Human-in-the-Loop

Automated systems alone often cannot meet all the criteria necessary for real-world deployment, such as ensuring reliability under unexpected conditions, transparency and accountability. This makes integrating human decision-makers with algorithmic systems a central concern in the public sector, where systems inform high-stakes decisions and need to comply with complex regulatory frameworks. Mitrou, Janssen, and Loukis (2022) highlight the need for human discretion and oversight when systems continuously learn on (biased) historical data,

face competing objectives and values, or need to meet accountability obligations. These concerns are often reflected in legal frameworks like the EU's proposal for AI regulation, which stresses the importance of human oversight for high-risk systems in Article 14 (EU Commission, 2021). Thus, the goal is not to replace humans with ADM systems but to assist public decision-makers in their tasks (Enarsson, Enqvist, & Naarttijärvi, 2022).

In such scenarios, humans must maintain the final say, making it important to consider how they interpret model outputs and integrate them into their decision-making process. This shifts the focus from simply building the most accurate prediction models to evaluating the consequences of providing specific model recommendations to human decision-makers (Fernández-Loría & Provost, 2022b; Vodrahalli, Gerstenberg, & Zou, 2022). This introduces new challenges for model developers, who need to ensure that the output of an ADM system can be effectively used by a human decision-maker.

Studies have shown that users are often hesitant to follow recommendations of a predictive model, a phenomenon known as algorithm aversion (De-Arteaga, Fogliato, & Chouldechova, 2020; Dietvorst, Simmons, & Massey, 2018, Dietvorst, Simmons, & Massey, 2015). This lies in contrast with the opposing tendency of automation bias, when humans excessively rely on a machine's suggestion (De-Arteaga et al., 2020; Goddard, Roudsari, & Wyatt, 2012). Ongoing research into how humans interact with algorithmic decision-making systems (Chugunova & Sele, 2023) highlights how these challenges differ based on application contexts and user characteristics. For instance, Cheng and Chouldechova (2022) demonstrate how less experienced child welfare hotline call workers tend to rely more on an algorithmic risk score than senior workers. Such insights need to guide model development to ensure that the technical design aligns with user requirements and preferences, enabling human decision-makers to make optimal choices. This can involve complex tradeoffs; for example, Chugunova and Sele (2023) illustrate that allowing users to modify algorithmic recommendations increases their willingness to adopt them, but tends to decrease decision accuracy.

For the interaction between human decision-makers and ML models to work, model predictions and its functioning must be comprehensible for human users (Amarasinghe et al., 2023; Nourani, Kabir, Mohseni, & Ragan, 2019; Yeomans, Shah, Mullainathan, & Kleinberg, 2019). For instance, Lebovitz, Lifshitz-Assaf, and Levina (2022) show how opaque ML models make it more difficult for medical professional to effectively use them for diagnosis. Many methods have been proposed to make ML models interpretable and explainable, with comprehensive overviews available in Belle and Papantonis (2021); Molnar (2022); Doshi-Velez and Kim (2017); Murdoch et al. (2019); Rudin et al. (2022). One of the reasons why these approaches vary widely is because they have very different conceptions of what constitutes an understandable explanation of the output of a model. Amarasinghe et al. (2023) establish an initial taxonomy linking public policy use cases with existing explainable ML approaches. Moving forward, they stress the need to rigorously evaluate explainable ML methods in real-world problem contexts to ensure their effectiveness in achieving real policy goals and in aiding domain experts. Research from fields such as psychology, cognitive sciences and philosophy may help in the task of creating explanations that are helpful to human users (Miller, 2019). This requires careful investigation of various challenges, such as identifying situations where model explanations may be harmful due to information overload (Poursabzi-Sangdeh, Goldstein, Hofman, Wortman Vaughan, & Wallach, 2021), or when users may take advantage of increased transparency to exploit a system (Molnar, 2022). Similarly, misleading explanations can be used to manipulate users and unjustifiably increase trust in a system (Lakkaraju & Bastani, 2020).

Providing uncertainty estimates for individual predictions can be critical for enhancing human decision-making based on algorithmic recommendations (Bhatt et al., 2021). Uncertainty estimates allow human decision-makers to assess the reliability of a prediction, and when it is necessary to manually intervene (Gruber et al., 2023; Shalit, 2022). This is particularly relevant due to the human tendency to rapidly lose trust in algorithmic systems upon observing errors, despite the algorithm's superior overall performance compared to human decision-makers (Dietvorst et al., 2015). Transparently communicating uncertainty to model users can therefore be a key element for building trust, which also illustrates the need for research into effective communication of probabilities to humans (Bhatt et al., 2021; Vodrahalli et al., 2022).

3. Expanding the ADM toolkit: Choosing the target estimand

In Section 2, we outlined several challenges that could threaten the intended functioning of an ADM system using supervised ML models to inform public sector decision-making. These challenges highlight several limitations of solely relying on a traditional supervised ML framework in model design (Wang et al., 2023). In response to these limitations, there have been calls to move beyond purely predictive modeling towards ML methodology that more directly centers around the goal of decision-making (Hüllermeier, 2021). This involves a shift in perspective from solely focusing on achieving accurate predictions to a more holistic modus operandi centered on selecting a modeling approach that can best inform the decision-making for a given policy goal and application context.

To illustrate this shift, we will discuss three distinct modeling frameworks, starting with standard risk prediction, which is commonly used in ADM systems, and then move to explore two additional causal modeling frameworks. Standard (risk) prediction (Section 3.1) focuses on estimating outcomes based on historical data without explicitly considering causality. Counterfactual modeling (Section 3.2) extends this approach by estimating causal outcomes of different hypothetical decisions, directly addressing issues such as the influence of past decision-making on the available outcome data. Lastly, policy learning (Section 3.3) aims to directly learn decision policies that maximize a predefined overarching utility, offering a practical approach to optimize a decision policy within the constraints of real-world scenarios (Section 2.4). Fig. 2 visually compares how well counterfactual prediction and policy learning address the challenges we have discussed in the context of standard prediction.

First, our goal is to examine the implicit and explicit assumptions underlying each modeling framework. This involves addressing two questions: 1) whether the target estimand in each approach is sufficiently linked to the decision-making process, meaning it would genuinely aid in making informed decisions. Understanding these connections is complex, as the guiding principles of an ADM system can be ambiguous, even when specific goals are in place. For example, public employment agencies often seek to allocate resources to job seekers at higher risk of long-term unemployment. However, this objective might stem from either a belief in the effectiveness of early interventions for high-risk job seekers or the notion that high-risk job seekers inherently deserve more support (Desiere & Struyven, 2021). Such ambiguity can present difficulties, complicating the choice of the appropriate target estimand, as a need-based distribution necessitates different modeling considerations than an approach focused on the most efficient allocation of interventions. 2) whether estimation is feasible, and what external assumptions are necessary to ensure the accuracy of such estimates. We will discuss these questions for each framework, allowing decision-makers to assess the validity of each approach for their application context.

We will then discuss how the (remaining) challenges outlined in Section 2 can be addressed within each framework. We will present methodological advancements specific to each modeling approach that can help overcome the discussed challenges. While some challenges may be common across all frameworks, others might be more pronounced in specific modeling approaches. Specifically, we focus on distribution shifts to ensure robustness across deployment environments, uncertainty quantification as a key building block for generating trustworthy

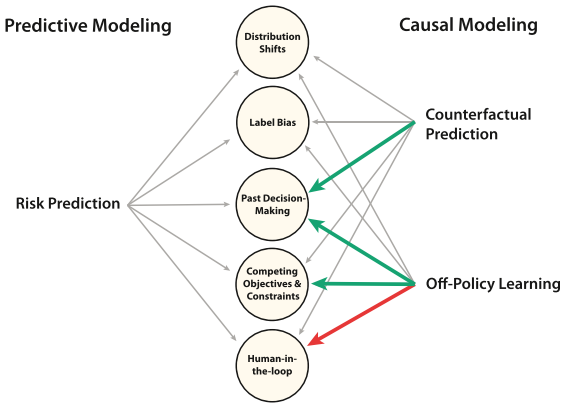


Fig. 2. Comparing ML Frameworks for Public Sector Decision-Making. Green lines indicate where a causal ML framework is potentially better at addressing a challenge, red lines highlight additional difficulties, and gray lines represent the baseline difficulty of using standard predictive modeling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predictions for humans, and multi-objective optimization to manage tradeoffs between competing objectives.

While we highlight three central modeling approaches, it is important to note that we do not address modeling approaches for every potential decision-making setting. Scenarios involving continuous interventions, interventions across time, sequential data, or interventions that simultaneously target multiple outcomes require other specialized approaches, which are beyond the scope of this paper (Acharki, Lugo, Bertoncello, & Garnier, 2023; Hüllermeier, 2021; Lin, Sperrin, Jenkins, Martin, & Peek, 2021; Van Geloven et al., 2020).

3.1. Risk prediction

In practice, ADM systems commonly involve optimizing predictive models for the estimation of individual outcomes used as decision-criteria (Wang et al., 2023). While predictive ML models are relatively straightforward to set up and train compared to causal models, relying on predictions as proxies for causal outcomes in decision-making runs the risk of generating misleading recommendations (Athey, 2017; Coston et al., 2020; Van Geloven et al., 2020; Wang et al., 2023). Nevertheless, in certain scenarios, risk predictions may still serve as a useful proxy for decision-making (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015; Guerdan, Coston, Wu, & Holstein, 2023; Fernández-Loría & Provost, 2022a). This is the case if the prediction target is still helpful with regards to the chosen allocation principle. For example, if the objective is to intervene only in the top-k of individuals, and the ranking induced by the non-causal predictions aligns with that of the causal outcomes, an accurate estimate of the intervention effects may not be critical (Fernández-Loría & Provost, 2022a).

The validity of such an approach hinges on external assumptions about the relationship between prediction proxies and causal effects. For example, if the predicted outcome is unaffected by interventions, but remains correlated with the causal effects, it can provide valuable information for the allocation strategy, even if it does not correspond directly to the target quantity being optimized. For instance, Kleinberg et al. (2015) discuss how predicting the mortality risk of patients during the next 1–12 months can be a helpful proxy variable for deciding which patients should not undergo hip and knee replacement surgeries. They argue that patients at risk of death during the months after surgery would not live long enough for the benefits of the surgery to outweigh its costs. Additionally, they assume that the surgery will not significantly impact the mortality risk after the first month, making it possible to determine the optimal intervention — whether to exclude a patient from the surgery or not — based on the predicted risk alone (Kleinberg et al., 2015).

However, settings in which a practical proxy for the intervention outcome exists may not be common in practice (Wang et al., 2023), because the connection between predicted risk scores and causal effects is rarely straightforward. For example, the predicted risk of death alone would not be sufficient to decide which patients should be first considered for surgery, because the benefits and potential complications of the treatment will probably vary among patients with the same risk score (Athey, 2017; Wang et al., 2023). Similarly, the Austrian Public Employment Services used a predictive model to assess the risk of long-term unemployment among job seekers with the goal of allocating interventions on this basis (Allhutter et al., 2020). This model divided job seekers into three risk groups. Medium risk individuals were prioritized for support, while high and low-risk individuals were given limited access to labor market programs. However, relying on risk scores to determine an efficient intervention assignment is questionable, as the effectiveness of labor market programs often varies among individuals (Cockx, Lechner, & Bollens, 2023), even those with the same risk score.

Risk prediction is centered in standard supervised ML methodology, aiming to estimate the statistical relationship between individual covariates X and outcomes Y by learning a prediction function $f : \mathscr{X} \to \mathscr{Y}$ from a set of observed training data $\mathscr{D} = \{(X_i, Y_i)\}_{i=1}^{n}$. Even assuming that these predictions provide useful information for the decision-making process, several general threats to the validity of using such a model, as discussed in the previous sections, remain. In the following sections, we will explore methods relevant for describing and tackling these challenges within the realm of risk prediction. This discussion will also lay the groundwork for addressing these challenges in the contexts of counterfactual prediction and policy learning.

3.1.1. Distribution Shifts, Selection Bias and Label Bias

Most supervised learning models assume that training and deployment data follow the same distribution. However, in many real-world scenarios we may encounter a distribution shift between the training

and deployment environment, necessitating the development of reliable models capable of handling and mitigating such differences (Duchi & Namkoong, 2021). Training models that remain valid under distribution shift often requires assumptions about the expected type of shift (David, Lu, Luu, & Pal, 2010). As outlined in Section 2.1, we will predominantly focus on covariate, label and concept shits (Moreno-Torres et al., 2012; Quinonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2008). Additionally, we will explore label bias as a type of distribution shift introduced through the use of proxy variables, and consider shifts in time caused by non-stationary environments. We outline central research streams below – for more in-depth introductions to transfer learning, domain adaptation and out-of-distribution generalization approaches, see Kouw and Loog (2018) and Zhou, Liu, Qiao, Xiang, and Loy (2022).

The survey research literature is an invaluable resource for understanding and systematizing different error sources in the data collection process that may surface as distribution shifts downstream. With their inherent focus on valid population inference, concepts such as undercoverage (relevant subpopulations cannot be reached with a data collection schema) and non-response (potentially selective non-participation of relevant subpopulations) extend beyond the traditional survey setting and can help to systematize deficits in the training data in relation to the target population. Error frameworks such as the Total Survey Error (Biemer, 2010; Groves & Lyberg, 2010) and its extensions (Sen, Flöck, Weller, Weiß, & Wagner, 2021) have been proposed to systematically trace errors along the data collection and processing pipeline that can accumulate in misrepresentation issues. Related work proposes strategies for improving inference from data that do not adequately represent the target population of interest (Cornesse et al., 2020; Yang & Kim, 2020). In this context, pseudo-weighting approaches (Elliott & Valliant, 2017; Valliant & Dever, 2011) are employed to match the potentially biased source data to some known reference distribution, which resembles methodology from the domain adaptation literature (see below) and similarly connects to concepts in causal inference (Mercer, Kreuter, Keeter, & Stuart, 2017). As a recent example of cross-disciplinary work in this context, Kim, Kern, Goldwasser, Kreuter, and Reingold (2022) draw on the multicalibration framework from algorithmic fairness (Hebert-Johnson, Kim, Reingold, & Rothblum, 2018) to learn prediction functions that are universally adaptable to unknown deployment shifts.

Domain adaptation techniques in the ML literature aim to construct a model that performs well in a setting different from but related to the one it was trained on (Hedegaard, Sheikh-Omar, & Iosifidis, 2021; Kouw & Loog, 2018). Unsupervised domain adaptation methods only make use of unlabeled target data to adjust the training data so it better aligns with the deployment distribution (Shimodaira, 2000; Subbaswamy, Chen, & Saria, 2022). For example, when a clinical risk prediction tool is deployed in a new hospital, a complete dataset may not be available to re-train the model for the new location. However, it might still be possible to adjust for potential covariate or label shift using unlabeled patient data, assuming that the underlying mechanisms between covariates and outcomes remain invariant. For example, the relationship between diseases and symptoms would not be expected to change between hospitals (Lipton, Wang, & Smola, 2018).

Given such data many domain adaptation methods involve importance-weighting, which makes use of the density ratio $w(X) = p_{\mathcal{T}}(X)/p_{\mathcal{S}}(X)$ or class proportions $w(Y) = p_{\mathcal{T}}(Y)/p_{\mathcal{S}}(Y)$ to adjust the loss function (Kouw & Loog, 2018; Shimodaira, 2000). Estimating these weights makes it possible to express the target risk relative to the source distribution $R_{\mathcal{T}}(L) = \mathbb{E}_{\mathcal{T}}[L(X,Y)] = \mathbb{E}_{\mathcal{S}}[w(X,Y)L(X,Y)]$ which then can be minimized (Fang, Lu, Niu, & Sugiyama, 2020). Various strategies can be used to estimate the importance weights, such as logistic regression (Bickel, Brückner, & Scheffer, 2009), kernel density estimation (Yu & Szepesvári, 2012), kernel mean matching (Quiñonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009) and KL-divergence minimization (Sugiyama, Nakajima, Kashima, Buenau, & Kawanabe, 2007).

However, weighting methods struggle in settings with limited and complex source data, frequently resulting in high variance estimates, and depend on data being available from the target domain of interest (Fang et al., 2020; Kouw & Loog, 2018; Liu & Ziebart, 2014). Distributionally robust methods offer an alternative approach by providing worst-case guarantees. They often involve minimax estimation, which seek to minimize the loss under the least favorable distribution shift (Duchi & Namkoong, 2021; Kouw & Loog, 2018; Subbaswamy et al., 2022; Wen, Yu, & Greiner, 2014).

In addition to the distribution shifts discussed so far, label bias and label noise can present significant challenges, especially given the prevalent use of proxy labels for decision-making. This bias arises when a model is trained not on the true latent label Y of interest but on an erroneous proxy label $\widetilde{Y}$. The label bias quantifies the difference between the true distribution of interest $p_T(Y|X)$, and the distribution $p_S(\widetilde{Y}|X)$ estimated from the proxy labels (Gruber et al., 2023). This shift can be characterized by a label corruption process or measurement error model, which describes the probability of a true label Y being recorded as a proxy label $\widetilde{Y}$ (Dai & Brown, 2020; Fang et al., 2020; Gruber et al., 2023).

Various approaches have been devised to mitigate label noise and measurement error. For instance, Natarajan, Dhillon, Ravikumar, and Tewari (2013) propose an unbiased risk minimization strategy for handling class-conditional $p(\widetilde{Y}|Y)$ noise. While such a simplified model of a proxy may be applicable in some settings, practitioners will likely encounter more complex scenarios (Chen, Ye, Chen, Zhao, & Heng, 2021), such as the measurement error depending on sensitive covariates (Obermeyer et al., 2019; Wang, Liu, & Levy, 2021). In some situations, there may be the option to access multiple proxies of the true target of interest. For example, Boeschoten, van Kesteren, Bagheri, & Oberski (2021) utilize a structural equation model to characterize the relationship between multiple proxies and the unobserved outcome to ensure fair predictions. There has also been research into how label bias interacts with other distribution shifts. For example, Dai and Brown (2020) propose a joint framework for addressing label bias and label shift, while Yu et al. (2020) investigate the interaction between class-conditional noise and generalized target shifts.

Distribution shifts tend to occur gradually over time (Webb, Hyde, Cao, Nguyen, & Petitjean, 2016), often triggered by the deployment of the model itself. Addressing such feedback loops and ongoing distribution shifts poses a significant challenge, likely requiring future research into the temporal dynamics of ML-informed decision-making (Pagan et al., 2023). For example, Perdomo et al. (2020) introduce a modeling framework that incorporates the potential impact of predictions on the predicted outcome of interest. These predictions are referred to as *performative*, effectively leading to distribution shifts by altering the target distribution in the deployment environment over time. They develop the notion of performative optimality, ensuring that a decision rule minimizes the expected loss with regard to the future target distribution it induces. The specific choice of the loss function can align with different objectives. For instance, one may opt to optimize for a target distribution with mostly favorable outcomes instead of solely focusing on accurate predictions (Kim & Perdomo, 2023).

3.1.2. Uncertainty Quantification

Accurate uncertainty estimates are key for enabling reliable decision-making systems. For example, they make it possible to determine when a model should refrain from making a recommendation and instead fully defer to a human user (Gruber et al., 2023). While we highlight selected methods below, we refer the reader to (Bhatt et al., 2021; Gruber et al., 2023; Hüllermeier & Waegeman, 2021; Sullivan, 2015) for comprehensive reviews of the emerging literature on uncertainty estimation in machine learning.

In recent years, interest has grown in conformal prediction as a distribution-free and model-agnostic approach to uncertainty

quantification for ML models. These characteristics make conformal prediction particularly appealing in many practical scenarios, as no specific assumptions on the model are required, enabling easy implementation for any arbitrary ML model. Instead of providing a point prediction, conformal prediction constructs a set of plausible predictions with respect to a chosen significance level (Angelopoulos & Bates, 2022; Papadopoulos, Proedrou, Vovk, & Gammerman, 2002a; Vovk, Gammerman, & Shafer, 2022). A larger conformal set indicates higher uncertainty in the model's predictions. Conformal prediction requires splitting the data into a training set and an additional holdout dataset, known as the calibration set. Alternatively, full conformal prediction does not necessitate dividing the data but is usually computationally more demanding (Angelopoulos & Bates, 2022). Conformal prediction relies on exchangeable data, which can not be guaranteed in scenarios involving distribution shifts. However, efforts have been made to extend conformal prediction for such situations, such as making use of weighting methods akin to those discussed in the context of unsupervised domain adaptation (Barber, Candès, Ramdas, & Tibshirani, 2023; Gibbs & Candes, 2021; Tibshirani, Foygel Barber, Candes, & Ramdas, 2019).

As mentioned, uncertainty estimates play a significant role in facilitating cooperative interaction between human users and models, particularly for high-stakes decision-making prevalent in the public sector. For example, Straitouri & Rodriguez (2024) propose a decision-support framework that makes use of conformal prediction to improve the cooperation between experts and the ML model. Their modeling framework restricts domain experts to choose their prediction from a set of plausible predictions generated by the model, resulting in better performance than relying on the model or the human expert alone.

3.1.3. Multi-Objective Optimization

A central challenge for ADM systems is balancing multiple objectives within a constrained outcome space. This often requires making tradeoffs between competing objectives, such as determining the appropriate balance between an equitable distribution of resources and maximizing cost-efficiency. Consequently, they require stakeholder input, further complicating the task by necessitating systems that are sufficiently accessible and interpretable for stakeholders to both make and evaluate these tradeoffs effectively (Papalexopoulos et al., 2022).

In the context of risk prediction, predictive modeling and decision-making are separated into two distinct steps (Elmachtoub & Grigas, 2022; Kuppler et al., 2022). Initially, a prediction is generated that subsequently gets used to inform a downstream allocation problem. In current ADM systems practice, multi-objective optimization is rarely employed. Typically, problems are cast as single-objective constrained optimization tasks, like finding the best allocation within budget constraints. When more complex constraints, different decision criteria and multiple predictions come into play, a conventional approach for formalizing the decision step involves the creation of a scalar utility function that linearly combines different objectives into a weighted sum (Keeney & Raiffa, 1993). For instance, stakeholders might construct a unified risk score out of multiple criteria that is subsequently employed to prioritize the allocation of resources. However, constructing a joint utility function can be tricky (Das & Dennis, 1997), as stakeholders often struggle to determine how to exactly weigh different objectives (Boutilier, 2013; Hayes et al., 2022; Roijers, Vamplew, Whiteson, & Dazeley, 2013). This difficulty is especially pronounced in risk prediction, where the link between predictions and the expected utility is often not entirely specified. A predicted risk score may only allow for a prioritization of individuals while the exact size of the individual utilities remains unknown. For example, in scenarios where intervention costs vary significantly by individual it might be important to compare the exact magnitude of the guiding utility for each individual intervention, making it problematic if only a ranked list is available.

A common alternative to defining a utility function a priori is to seek allocations that reside along the Pareto front, which constitutes the set solutions where improving one objective necessarily entails the worsening of another (Deb, 2011; Hayes et al., 2022). For example, Hertweck, Baumann, Loi, Viganò, and Heitz (2022) propose a framework to visualize tradeoffs between the utility of the decision-maker and the fairness demands of the decision subject. While approaches like this will still leave stakeholders with difficult value choices, they might aid in making tradeoffs more explicit by focusing the selection on a specific set of allocation policies. Similarly, the notion of multi-target multiplicity describes a scenario in which multiple prediction targets that are all considered to be equally valid operationalizations of the outcome of interest are available (Watson-Daniels, Barocas, Hofman, & Chouldechova, 2023). This makes it possible to explore arbitrary combinations of these targets to arrive at an allocation that maximizes group-level fairness.

On the other hand, secondary objectives and constraints such as ensuring models are fair (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Hardt, Price, Price, & Srebro, 2016; Hort, Chen, Zhang, Harman and Sarro, 2023; Zafar, Valera, Rogriguez, & Gummadi, 2017) and interpretable (Molnar, 2022) might already come into play in the modeling process. More efforts are being made to naturally integrate such constraints into the ML pipeline. For example, recent work in Multi-Criteria Auto ML (Pfisterer, Coors, Thomas, & Bischl, 2019) proposes a framework where users can iteratively specify tradeoffs between different objectives, such as fairness, accuracy and robustness, to explore subregions of the Pareto front. Automatized modeling procedures of this kind might make it easier to interactively elicit stakeholder preferences.

3.2. Counterfactual prediction

The primary goal of any ADM system is to guide decision-making by recommending a particular course of action. Making such recommendations effectively will often involve counterfactual modeling. While non-causal risk predictions can be used as proxies for relevant counterfactual outcomes, they risk being significantly biased, potentially making a more principled approach involving explicit causal modeling preferable. However, as outlined in Section 2.3, a common threat to the validity of causal models is confounding, requiring external assumptions and historical data on intervention assignment to address. This challenge requires careful case-by-case analysis by the model developer and may limit the possibility to make use of counterfactual estimates for decision-making in certain application contexts.

The potential outcomes framework (Rubin, 1974) is a prominent approach for framing causal questions. In a binary intervention scenario $\mathcal{T} = \{0, 1\}$, it denotes two potential outcomes $(Y_i(0), Y_i(1))$ for an individual i. These outcomes represent the two possible observable outcomes: no intervention ($T_i = 0$) and an intervention ($T_i = 1$). The individual treatment effect τ_i is then defined as the difference between the potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. Estimating potential outcomes and treatment effects from observational data $\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^{n}$ is challenging, as it is usually only possible to observe one outcome $Y_i = (1 - T_i)Y_i(0) + T_i Y_i(1)$ for each individual (Künzel, Sekhon, Bickel, & Yu, 2019). Consequently, it is common to estimate the expected potential outcomes $\mu_t(x) = \mathbb{E}[Y(t)\,|X = x]$ and conditional average treatment effect (CATE) $\tau(x) = \mathbb{E}[Y(1) - Y(0)\,|X = x]$ for a given covariate vector $X = x$ (Künzel et al., 2019; Vegetabile, 2021). We refer to Appendix B for an overview of relevant ML-based CATE estimation methods.

To link the CATE with a statistical estimand, a set of untestable assumptions is required (Caron et al., 2022; Johansson, Shalit, Kallus, & Sontag, 2022; Künzel et al., 2019). Unconfoundedness $(Y(0), Y(1)) \perp T|X$, requires that potential outcomes are conditionally independent of treatment assignment. Positivity guarantees nonzero propensity scores $0 < P(T = 1|X = x) < 1$ for all confounders $x \in \mathcal{X}$, meaning that treatment assignment is not fully deterministic. Finally, Stable Unit

Treatment Value Assumption (SUTVA) assumes that the outcome of one individual is not affected by the interventions others received, and that there are no different versions of a specific treatment. Under these assumptions it becomes in principle possible to infer $\mu_t(x)$ and $\tau(x)$ from observational data (Caron et al., 2022). Ensuring these assumptions can be difficult, with unmeasured confounding posing a significant risk when aiming for valid counterfactual predictions.

While the individual treatment effect seems like a natural choice for determining the optimal allocation, there exist various scenarios in which estimating only one expected potential outcome $\mu_t(x)$ may be sufficient to inform the decision-making process (Dickerman & Hernán, 2020). These outcomes could, for example, represent the likelihood of abuse if a hotline call is not followed up (Coston et al., 2020) or the risk of death of a patient if no heart transplant is performed (Dickerman & Hernán, 2020; Van Geloven et al., 2020). Models that provide such risk assessments align well with allocation principles informed by need-based criteria by identifying individuals at high risk of adverse outcomes. For example, when screening phone calls for potential child maltreatment (Vaithianathan et al., 2019), there exists a moral and legal obligation to investigate high risk cases, regardless of the investigation's likelihood of success (Chouldechova et al., 2018; Coston et al., 2020). Additionally, in scenarios where one potential outcome is trivially known, only one outcome needs to be estimated. For instance, in judicial bail prediction, individuals for whom bail was denied cannot re-offend before trial (Lakkaraju et al., 2017).

While assumptions for causal identification are still necessary to correctly estimate expected potential outcomes, in many scenarios this task may be more feasible than full treatment effect estimation. For example, this might be the case when implementing an intervention that has not been previously deployed, or when data on specific outcomes is generally limited (Fernández-Loría & Provost, 2022a). Following the discussion on proxies in risk prediction, explicitly modeling the treatment effect might also not be necessary if the relationship between a potential outcome and treatment effect is well-established (Fernández-Loría & Provost, 2022a). For example, prior knowledge and experiments may indicate that a particular treatment strategy is the most beneficial approach for individuals in a specific risk group (Athey, Keleher, & Spiess, 2023), allowing us to correctly prioritize individuals based on the estimated baseline risk $Y(0)$ alone (Fernández-Loría & Loría, 2023).

Compared to CATE estimation, this only requires the estimation of a single outcome regression $\mu_t(x) = \mathbb{E}[Y|X = x, T = t]$. While this simplifies some of the necessary considerations for CATE estimation, caution may still be warranted in low-data settings due to differences in the covariate distribution of the treatment group and overall population, potentially necessitating dedicated approaches to correct this imbalance during estimation (Johansson et al., 2022). A growing body of recent research focuses on auditing and evaluating counterfactual prediction models of this nature for algorithmic decision-making. For example, Coston et al. (2020) discuss evaluation fairness metrics and evaluation methods for counterfactual risk modeling. In the domain of clinical risk prediction, a substantial body of literature explores methods for predicting outcomes under specific medical treatments (Lin et al., 2021; Prosperi et al., 2020; Schulam & Saria, 2017; Van Geloven et al., 2020).

In the following, we will discuss approaches to address distribution shifts, uncertainty quantification and multi-objective optimization for CATE estimation. While many of the earlier considerations in the context of risk prediction remain applicable, there are aspects unique to this setting, requiring methods dedicated to tackling the challenges for causal modeling.

3.2.1. Distribution Shifts, Selection Bias and Label Bias

The challenge of handling distribution shifts is strongly related to causal estimation, as illustrated by Johansson, Shalit, and Sontag (2016). For example, predicting counterfactual outcomes under no unmeasured confounding corresponds to unsupervised domain adaptation under covariate shift (Johansson et al., 2022). This is because past decision-making policies often lead to a difference in covariate distribution between the treatment groups and the distribution of the overall population. Several approaches for dealing with shifts when performing CATE estimation have been proposed (Assaad et al., 2021; Johansson et al., 2016; Shalit, Johansson, & Sontag, 2017). For example, Kuzmanovic et al. (2023) study the problem of inferring CATE in settings in which treatment information is missing for some individuals, a challenge they frame as a covariate shift problem.

In many practical settings, approaches geared towards guaranteeing robustness to unknown distribution shifts (Jeong & Namkoong, 2020) may be particularly relevant, as it can be difficult to anticipate the target population and relevant subpopulations in all possible deployment environments. To tackle this challenge, Kern, Kim, and Zhou (2024) introduce an approach for learning robust CATE estimates under unknown external covariate shifts. They achieve this by employing a boosting-style post-processing routine to generate a multi-accurate predictor (Kim, Ghorbani, & Zou, 2019), enabling unbiased predictions in a new deployment setting.

In a recent study, Guerdan, Coston, Wu, and Holstein (2023) examine the interaction of label bias and counterfactual prediction. They propose a causal framework that describes potential biases introduced by proxy labels, and survey strategies for evaluating the chosen measurement model. There are not many approaches that explicitly deal with measurement error in the context of employing counterfactual models. Guerdan, Coston, Holstein, and Wu (2023) develop a framework that accounts for treatment-conditional errors based on the previously discussed approach for correcting class-conditional noise (Natarajan et al., 2013).

3.2.2. Uncertainty Quantification

Recently, conformal prediction has been extended to address individual treatment effect estimation, with a central challenge being that exchangeability of the data can not be guaranteed due the covariate shift between treatment groups and the overall population (Alaa, Ahmad, & van der Laan, 2023). Lei and Candès (2021) propose a solution that makes use of weighted conformal prediction (Tibshirani et al., 2019) to correct for this shift. They construct prediction intervals for potential outcomes, which are then used to derive intervals for the individual treatment effects. In contrast, conformal meta-learners, as introduced by (Alaa et al., 2023), offer a framework for directly constructing prediction intervals for pseudo-outcomes of two-stage meta-learners, allowing for conformal prediction for a different class of CATE estimation methods. As in the case of risk prediction, providing a conformal set has the potential to facilitate human and model interaction by guiding a decision-maker towards a set of likely solutions, while still leaving the critical final decision to the human.

3.2.3. Multi-Objective Optimization

In principle, the challenge of handling multiple objectives and constraints remains similar when employing risk prediction and counterfactual prediction. In both scenarios, multi-objective optimization typically becomes relevant when determining the downstream allocation after a prediction is generated. However, counterfactual estimates are usually easier to link with the intended guiding utility than non-causal predictions, making it more straightforward to quantify trade-offs with other objectives.

Efforts have been made to explicitly integrate CATE estimation and prescriptive optimization within an unified framework, typically with a focus on budget-constrained optimization problems (Ai et al., 2022; Tu et al., 2021). Formulating such optimization problems is generally made easier when the expected net benefit can be easily defined, such as maximizing net revenue when allocating tax audits within a fixed budget (Black et al., 2022). Similarly, McFowland III, Gangarapu, Bapna, and Sun (2021) present a prescriptive analytics framework that combines randomized experiments, CATE estimation and a subsequent

4 Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

constrained optimization problem to identify the profit-maximizing allocation policy. Crucially, the expected cost of an intervention may not necessarily be known, potentially requiring a separate estimation process. Unlike in the case of generic prediction, only a few studies have attempted to integrate constraints directly into the counterfactual estimation process. Notable examples include Kim and Zubizarreta (2023) for CATE estimation and Mishler et al. (2021) for counterfactual risk prediction under fairness constraints.

3.3. Policy learning

The ADM approaches discussed so far entail a two-step process: initially estimating individual outcomes, such as the CATE, and subsequently using these estimates to determine an optimal downstream allocation considering external constraints. However, this means that the target of estimation is only indirectly linked to the underlying policy objective, as an improved prediction may not necessarily enhance the utility of the resulting allocation policy (Perdomo, 2024). While perfect predictions could in theory lead to optimal decision-making, in practice it may sometimes more practical to estimate the allocation policy directly (Elmachtoub & Grigas, 2022; Fernández-Loría & Provost, 2022a). A wide range of methods have been proposed to optimize the aggregated utility, emphasizing that the primary goal of deploying a statistical targeting system is not accurate prediction alone.

More specifically, the target of estimation becomes the allocation policy $\pi : \mathscr{X} \to \{0, 1\}$, directly mapping individual covariates X_i to the likelihood of intervening. Learning optimal assignment rules has been studied across different disciplines, such as statistical decision theory, economics and operations research (Elmachtoub & Grigas, 2022; Kitagawa & Tetenov, 2018; Manski, 2004). This paper specifically highlights recent literature focusing on learning optimal policies from past observational data using ML methods (Athey & Wager, 2021; Hatamyar & Kreif, 2023; Kallus, 2018; Luedtke & van der Laan, 2016). Here, we are concerned with off-policy learning, given that on-policy learning may not be suitable for high-stakes settings in the public sector where active experimentation with different decision policies is not possible.

Off-Policy learning usually involves optimizing over a class of policies $\pi \in \Pi$ by defining the aggregated utility of a proposed policy $V(\pi) = \mathbb{E}[Y(\pi(x))]$ as the overall expected outcome if the policy were deployed (Athey & Wager, 2021). Finding the optimal policy corresponds to identifying the policy that maximizes the utility, i.e. $\hat{\pi} = \textit{argmax}_{\pi \in \Pi} \hat{V}(\pi)$. When estimating the policy value from observational data, we rely on the same assumptions as those used in CATE estimation to ensure identification, such as unconfoundedness. We refer to Appendix C for an overview of off-policy learning methods.

Adopting a population-level perspective and directly optimizing for the best allocation policy can come with several benefits. First, such an approach may aid in the natural integration of downstream constraints, for example by limiting the class of policies Π under consideration to these that can feasibly be implemented. For instance, policy learning enables the exclusion of decision policies that make use of specific covariates (Athey & Wager, 2021; Kallus, 2021), such as sensitive attributes like race and gender, or features susceptible to individual manipulation potentially leading to distribution shifts after deployment. While these variables may be required to address confounding during estimation, we can ensure that the decision-making does not rely on them by constraining the class of allowed policies. Second, estimating and optimizing the policy value $V(\pi)$ for a constrained set of policies is a distinct and potentially easier estimation task compared to predicting individual-level treatment effects (Kallus, 2021; Lechner, 2023). In general, precise estimation of individual-level outcomes may not always be a prerequisite for determining the optimal policy, as an erroneous prediction may not necessarily lead to an erroneous decision (Fernández-Loría & Provost, 2022a).

The feasibility of employing policy learning also depends on whether the goal of the modeling process is a fully automated decision system or providing recommendations to human decision-makers. As discussed in Section 2.5, fully formalizing the connection between model output and decision-making can be challenging due to the involvement of human decision-makers who may want to integrate external information and can overrule the model's recommendation (Shalit, 2022). This complication adds nuance to the discussion about choosing the most appropriate target estimand and modeling framework. For example, a human decision-maker might find a CATE estimate more trustworthy and more suitable for individual decision-making, as opposed to a fully defined allocation policy (Coston et al., 2020). Conversely, a well-defined policy class could also be restricted to policies that can be easily interpreted by users, such as decision trees (Athey & Wager, 2021).

In the next section, we will highlight relevant work extending policy learning to tackle distribution shifts, uncertainty quantification and multi-objective optimization. While policy learning shares many similarities with CATE estimation, it still involves a distinct target estimand and estimation strategies, requiring tailored approaches to this setting. Research at the intersections of policy learning and the aforementioned challenges is still in early stages, but there have been some promising developments in the recent past.

3.3.1. Distribution Shift, Selection Bias and Label Bias

As described, a significant challenge in managing distribution shifts for ADM systems lies in precisely specifying the anticipated changes from the historical environment to the future deployment environment. For example, the data-generating mechanism may change over time, but the exact nature of this shift is usually hard to predict. Addressing this challenge, Si, Zhang, Zhou, & Blanchet (2020, 2023) propose an algorithm for distributionally robust policy learning under unknown covariate and concept shift. Their approach involves maximizing the worst-case policy value over all environments within a specific distance to the training environment. By choosing their preferred distance, decision-makers can manage their risk aversion before deploying a policy (Si et al., 2023). Building on this work, Kallus, Mao, Wang, and Zhou (2022) incorporate doubly-robust methods, removing the need to assume that the historical assignment policy is known, which is often unavailable when relying on observational data. Instead of ensuring robustness under arbitrary distribution shifts, it may also be helpful to focus on specific types of shifts, potentially simplifying the integration of domain knowledge. For example, Hatt, Tschernutter, and Feuerriegel (2022) develop a framework for learning worst-case policies that generalize under distributional shifts resulting from an unknown selection bias.

3.3.2. Uncertainty Quantification

After selecting a policy for deployment, especially in high-stakes settings, reliable uncertainty estimates become important to guarantee the policy's reliability. Uncertainty quantification in off-policy learning often involves estimating bounds for the expected aggregate utility of the policy under hypothetical deployment (Taufiq, Ton, Cornish, Teh, & Doucet, 2022), for example as seen in Wang, Agarwal, and Dudík (2017). However, in many scenarios there may arise the need to quantify the uncertainty of outcomes at the individual level. For instance, a policy that appears to lead to a positive aggregate utility may still be deemed unacceptable if the variability in outcomes for certain subgroups is overly large. Recent works in off-policy evaluation have investigated the application of conformal prediction to construct prediction intervals. Similar to CATE estimation, a critical challenge for conformal off-policy prediction lies in guaranteeing exchangeability of the data. Zhang, Shi, and Luo (2023) and Taufiq et al. (2022) have proposed approaches that make use of weighted conformal prediction to address the shift between training data and deployment environment, allowing for the reliable estimation of prediction sets.

Table 1
Overview of different (causal) ML frameworks for Algorithmic Decision-Making. The validity of each approach is highly context-dependent, and requires careful evaluation of the available data, decision-making processes and policy objectives.

	Risk Prediction	Counterfactual Prediction		Off-Policy Learning
Estimand	$\mathbb{E}[Y\|X=x]$	$\mathbb{E}[Y(t)\|X=x]$	$\mathbb{E}[Y(1)-Y(0)\|X=x]$	$\hat{\pi}=\arg\max_{\pi\in\Pi}\hat{V}(\pi)$
Estimation & Data	$\{(X_i,Y_i)\}_{i=1}^n$ Y not influenced by interventions	$\{(X_i,Y_i,T_i=t)\}_{i=1}^n$ Causal identification and data for $T=t$	$\{(X_i,Y_i,T_i)\}_{i=1}^n$ Causal identification and data from both intervention groups	
Methods	Supervised ML methods	Supervised ML methods Re-weighting methods to correct covariate shift between intervention group and population	• Meta-Learners – Two-Stage Learners – Direct Estimators • Model-Specific Estimators	• Plug-in Estimators • IPW Estimation • DR Methods
Link to Policy Objective	Established link between non-causal prediction and utility of a decision	Outcome under intervention directly corresponds to objective, e.g. need-based allocation Prior information on relationship between outcome and effects, e.g. higher baseline risk implies a more effective intervention	Estimated effects allow for efficient allocation with regards to specified objective	Policy value encodes aggregated utility of a proposed allocation
Decision-Making	Predictions inform downstream decision-making process, e.g. recommendations for human decision-makers Multiple constraints and objectives usually considered after estimation			Finalized allocation policy for given utility and constraints Less suited for human-in-the-loop
Heuristic Overview	Least assumptions for estimation → Most assumptions for estimation Most assumptions for linking estimand to policy objective ← Least assumptions for linking estimand to policy objective			

4 Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

3.3.3. Multi-Objective Optimization

To generalize the policy value for multiple objectives, one can consider the weighted sum of utilities resulting from various individual outcomes and external objectives. For example, in scenarios with individually varying intervention costs, it may be possible to define a net-monetary benefit, that is subsequently used as the target outcome in the policy value (Xu et al., 2022). Alternatively, a decision-maker may want to enforce an overarching constraint, such a limited budget, as part of the policy optimization problem (Huang & Xu, 2020; Sun, 2021). However, if the relevant constraint needs to be adjusted regularly, such approaches may lead to significant computational costs (Sun, Munro, Kalashnov, Du, & Wager, 2024). Sun et al. (2024) propose learning an individual-level priority score that directly encodes the cost-benefit ratio, which can subsequently easily be used to rank individuals for intervention under varying resource constraints. Furthermore, off-policy learning methods that optimize within a constrained class of policies (Athey & Wager, 2021; Kallus, 2021) have the advantage that secondary constraints can also be encoded through restricting the class of allowed policies. For example, Frauen, Melnychuk, & Feuerriegel (2024) propose a policy learning that restricts the policy class to those respecting fairness constraints.

In practice, decision-makers frequently encounter scenarios with multiple objectives that are not easily expressed as constraints or monetary monetary costs. As described, identifying the set of Pareto-optimal models may allow stakeholders to effectively explore tradeoffs between objectives. Rehill and Biddle (2024) propose a multi-objective Bayesian optimization approach for off-policy learning, utilizing proxy models to efficiently construct the Pareto-Frontier, enabling a human user to better evaluate the consequences of different weightings between objectives.

4. Towards a decision-centric ML toolkit in the public sector

Making productive use of ML for public sector decision-making is a complicated task, requiring careful alignment of policy objectives and model development. First, we set out to explore challenges faced when deploying ML for public sector decision-making. Specifically, we focused on challenges arising from a misalignment between policy objectives and technical design, such as when assumptions about the training data do not match the intended application context. We identified and discussed five key challenges in Section 2, highlighting potential limitations of solely relying on the standard supervised ML paradigm (see Fig. 2). In response to these limitations, we examined alternative modeling frameworks, specifically counterfactual prediction and off-policy learning. Each framework comes with its own set of distinct advantages and is potentially better suited to overcome some of the challenges. To choose between these frameworks, a model developer should consider two key questions. 1) How will the estimated quantity be helpful in decision-making? 2) What is necessary to ensure unbiased estimation?

We observe that targets that are easier to estimate, often require more assumptions about their utility in the decision-making process. Even with perfect knowledge of these targets, they might need to be combined with domain knowledge to be informative for the decision-maker. Conversely, estimating causal outcomes requires more assumptions for causal identification, but might offer more actionable insight for decision-makers. We provide a distilled version of our presentation of different modeling approaches and their links to policy objectives and decision-making in Table 1, with the aim of providing guidance for discussions on which modeling theme is most suitable in a given scenario.

For example, traditional (risk) prediction methods, while requiring fewer assumption during the estimation process, are not well-suited to estimate counterfactual outcomes, which are often the true target of interest for decision-makers. This limitation often requires additional assumptions about how the predictions relate to the decision-maker's objective. Consider the previously discussed medical scenario as an example for such an assumption: the decision-maker assumes that a higher mortality risk in the coming years might make a knee replacement less beneficial, while also assuming that the mortality risk is not significantly impacted by the surgery itself. Counterfactual prediction can potentially support a variety of decision-makers goals with fewer explicit assumptions about how the goals of the decision-maker align with the target of estimation. For example, consider a public employment service aiming to target support measures to job seekers with a high risk of becoming long-term unemployed. Similarly, reliable estimates of the heterogeneous causal effects of a support program would aid a decision-maker in matching individual's to the most effective program. However, counterfactual estimation is more involved than standard (risk) prediction, relying on external assumptions to ensure causal identification. Policy learning is explicitly integrated into the decision-making process by optimizing a predefined utility function to directly estimate an allocation policy. Such an end-to-end approach allows for a more straightforward integration of constraints and additional objectives. However, it might struggle in scenarios in which human decision-makers play a crucial role. In such settings, decision-makers might prefer individual-level estimates as recommendations to guide their own judgments. We present three examples inspired by real-world use cases in Figs. A1, A2 and A3, each focusing on risk prediction, counterfactual prediction and policy learning respectively. They summarize some of the key questions practitioners need to address to ensure that the selected approach fits the intended application context.

Certain challenges such as the influence of past decision-making are inherently addressed by causal modeling frameworks. To tackle the remaining challenges, we have compiled for each modeling approach a selection of methods to address them, as detailed in the previous section and summarized in Table A1 in the appendix. Our goal was to identify methods that are applicable across various ML models within each respective modeling framework, to keep most of our discussion model agnostic and relevant across many application contexts. Unsurprisingly, there is generally less research addressing specific challenges within newer causal modeling frameworks. This gap presents a compelling direction for future research to explore which mitigation strategies from standard supervised ML could be extended to the causal setting.

In recent years, several studies have applied causal ML frameworks to practical applications in the public sector. For example, these modeling approaches have been used and evaluated for the optimal allocation of development aid (Kuzmanovic et al., 2024), allocation of medical preventive care (Kraus, Feuerriegel, & Saar-Tsechansky, 2024), child welfare hotline screening (Coston et al., 2020), and assignment of training programs to job seekers (Cockx et al., 2023).

5. Discussion

We analyzed challenges that result from a misalignment between the (technical) assumptions made during model design and the intended policy goals. Each challenge can lead to harmful unintended consequences, impacting the individuals affected by the decisions and potentially undermining the legitimacy of the system.

- *Distribution Shifts* occur when the data used to train the ML model does not reflect real-world conditions, causing the performance of the model to decline. Such shifts can lead to misclassifications of individuals and create harmful feedback loops that reinforce erroneous predictions.
- *Label Bias* can happen when a ML model relies on proxy variables to estimate hard to measure outcomes. If these proxies are biased and primarily reflect institutional practices instead of of the true target, they can systematically disadvantage certain groups, leading to unfair predictions and decision outcomes.
- *Past Decision-Making* often influences the available training data. If we do not explicitly account for these past interventions, the predictions of the ADM system may become outdated once new

Table 2
Overview of key challenges in ADM systems and technical solution strategies, focusing on the role of domain expertise in guiding technical design choices.

Challenge	Technical Solution Strategies	Stakeholder Input
Distribution Shifts	• Apply domain adaptation methods using data from the deployment environment • Use distributionally robust optimization for worst-case guarantees • Implement continuous monitoring of input data and model performance	• Collaborate with domain experts and data providers to anticipate changes in the deployment environment (e.g. changing regulations and user behavior) • Engage stakeholders to determine acceptable risk tolerance • Involve decision-makers to determine relevant evaluation metrics for ongoing monitoring • Identify vulnerable and hard-to-reach sub-populations
Label Bias	• Construct measurement error models for selected proxy variables • Validate chosen proxy variables using external data sources and additional variables	• Collaborate with domain experts to understand how proxy variables map to the true concepts of interest • Identify societal biases that may impact the proxy-target relationship
Past Decision-Making	• Identify suitable (causal) estimands and estimation strategies, such as CATE estimation, counterfactual prediction or policy learning	• Ensure that the chosen (causal) estimand is able to inform the decision-making process • Make use of domain expertise to inform assumptions necessary for causal identification, such as gathering knowledge on past decision-making criteria and processes
Competing Objectives and Constraints	• Integrate external constraints into model design and allocation procedure (e.g. using model multiplicty and constrained optimization) • Identify solutions along the Pareto frontier to enable decision-makers to manage tradeoffs	• Elicit preferences from decision-makers to quantify tradeoffs between different objectives and constraints • Collaborate with stakeholders to identify objectives not fully captured by the ADM system
Human-in-the-Loop	• Use uncertainty quantification (e.g. conformal prediction) and explainable ML methods to provide guarantees to decision-makers and enhance transparency	• Understand how model outputs will be interpreted and used by decision-makers, considering user background and workflows • Regularly gather feedback from end-users to improve model integration

decision-making practices are implemented. For example, a model might underestimate the risk for individuals who previously received support, potentially leading to harm if future allocations fail to consider this.

- *Competing Objectives and Constraints* can complicate the formalization of policy goals in an ADM system, as predictive systems often focus on singular, clearly defined objectives. This narrow focus can introduce omitted-payoff bias, such as when optimizing for cost-efficiency disproportionately harms marginalized groups.
- *Human-in-the-Loop* is an important component of ADM systems, because automated systems alone often do not meet all necessary requirements for real-world deployment. However, interactions between models and human decision-maker can introduce complications and biases. For example, an accurate prediction may not be helpful if case workers lack trust due to unclear communication of model uncertainties.

In this work, we found that standard machine learning approaches are not necessarily well-suited for the public sector context, requiring model designers to expand their toolkit to effectively address these issues. We discussed different technical solution strategies in detail and provided guidance on choosing between alternative modeling frameworks - including predictive and causal modeling - to tackle these challenges. However, to do so effectively, the technical model design needs to be guided in close collaboration with domain experts. Each challenge we presented is embedded in complex, changing social contexts, where purely technical solutions often fall short. The right technical design choice often has no clear or definite answer and can not be left to the model developer alone. For example, the implicit assumptions a model developer makes about the data distribution, the causal structure of the problem, or how different objectives should be represented in the system can greatly affect the validity of the resulting system. However, these assumptions often need to be informed domain-specific knowledge, necessitating insights from policy makers, social scientists and other stakeholders involved. In Table 2, we summarize technical solution strategies for each challenge and highlight the points where external input may be central to ensure that technical design choices remain aligned with real-world policy objectives. However, engaging and collaborating with stakeholders is often difficult in practice. Participatory design of AI systems is often mentioned as an important approach, but can be difficult to implement effectively. By specifying the points along the ML pipeline where stakeholders collaboration is crucial for supporting technical design decisions, we hope to guide these efforts and make participatory approaches more actionable (Delgado, Yang, Madaio, & Yang, 2023).

However, our focus does not cover all potential issues related to the use of predictive algorithms in the public sector. While ensuring that a system accurately reflects the goals of decision-makers is a prerequisite for ethical and reliable use, this alone is not sufficient. We do not explicitly discuss ethical, legal and broader societal challenges. Even if a system functions perfectly according to its intended goals, it can still lead to adverse outcomes. For example, this can happen if a decision-maker does not prioritize fair treatment of sensitive subgroups as an explicit design goal (Barocas et al., 2023).

The intention behind this paper was to clarify the assumptions behind different (predictive) modeling approaches, and help practitioners identify where common technical assumptions may not hold true in the public sector context. We see this as a first step towards the development of a robust methodological framework for constructing and maintaining ADM systems in the public sector that includes both best practices for practitioners and an up-to-date array of technical approaches. While we have made some inroads here by selecting and consolidating relevant theoretical advancements, a critical need for more research to connect these methods with real-world policy use cases remains. As illustrated by the methods and challenges presented here, achieving this goal will likely require bringing together researchers from various disciplines to develop systems that genuinely improve decision-making in tangible ways. It will require a shift in perspective away from technical approaches purely centered around predictive optimization and towards ones that explicitly incorporate decision-making and its impact into the modeling framework.

Effectuating this shift will involve opening up the ML pipeline to external input at significant points. On the one hand, this will require the development of effective strategies for engaging stakeholders and harnessing their expertise, particularly in integrating domain knowledge into the model-building process and eliciting values and objectives from decision-makers. At the same time, more work will have to be done to figure out how the development of ADM systems interacts with and can be embedded into institutional processes and structures. The prospect of this shift might seem daunting at first. It will involve establishing both technical and institutional frameworks that enable the development of successful ADM systems. However, this path also holds the potential to

transform our public policy processes in a positive way. It could lead to the establishment of new standards of transparency and the explication of previously implicit goals, as well as facilitating the development of new structures to integrate domain knowledge and involve important stakeholders. Thus, bridging the gap between explicit formalization and nuanced policy requirements could not only unlock the potential of successful ML applications in the public sector, but also lead to a public sector that is more understandable, open to scrutiny and thus accountable.

6. Conclusion

In this paper, we analyzed misalignments between the assumptions underlying ML models and the realities of public sector decision-making. We isolated and discussed five central challenges: distribution shifts, label bias, the influence of past decision-making, competing objectives and constraints and the integration of human decision-makers. We demonstrated how misalignment can lead to unreliable and harmful predictions, potentially causing systems to fail in achieving the intended policy goals and undermining the legitimacy of the decision-making process. Through our analysis, we concluded that many assumptions commonly made in the implementation of ML models do not hold in complex, evolving decision-making environments. In response, we argue for a shift in modeling efforts from focusing solely on predictive accuracy to improving decision-making outcomes. We presented alternative modeling approaches, including causal machine learning methods including counterfactual prediction and policy learning, which may be better suited to inform decision-making. We also provided guidance on selecting the appropriate modeling strategy by clarifying the assumptions underlying these approaches. Model developers should carefully consider how the estimated quantities can guide decision-making and whether unbiased estimation is possible given available data and external assumptions. Additionally, we summarized technical solutions to the discussed challenges, such as distributionally robust optimization, uncertainty quantification and multi-objective optimization within each modeling framework. Finally, we found that selecting the right methods and frameworks requires external input from domain experts and stakeholders to ensure that the implicit assumptions made by model developers align with the specific problem setting.

CRediT authorship contribution statement

Unai Fischer-Abaigar: Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Christoph Kern:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Noam Barda:** Writing – review & editing. **Frauke Kreuter:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial In- telligence, sponsored by the Federal Ministry of Education and Research, by the Baden-Württemberg Foundation under the grant "Fairness in Automated Decision-Making - FairADM" and by the Volkswagen Foundation, grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)". We express our gratitude to Nanina Föhr, Felix Henninger, Patrick Schenk, and Daria Szafran for their valuable feedback and support, and to Bernd Fitzenberger, for shaping our thinking about applications of prediction models in the public sector.

Appendix A. Methodological Approaches for ADM Systems in Public Sector Decision-Making

Table A1
Overview of Methodological Approaches to Address Key Challenges of ADM Systems in the Public Sector in Risk Prediction, Counterfactual Prediction, and Off-Policy Learning.

	Risk Prediction	Counterfactual Prediction	Off-Policy Learning
Distribution Shifts	• Biases in Data Collection (Biemer, 2010; Cornesse et al., 2020; Elliott & Valliant, 2017; Groves & Lyberg, 2010; Kim et al., 2022; Yang & Kim, 2020) • Domain Adaptation Using Data from Deployment Environment (Kouw & Loog, 2018; Shimodaira, 2000; Fang et al., 2020) • Worst-Case Guarantees (Duchi & Namkoong, 2021; Wen et al., 2014; Zhang et al., 2021) • Shifts induced by Model Predictions (Kim & Perdomo, 2023; Pagan et al., 2023; Perdomo et al., 2020)	• Covariate Shift between Intervention Groups (Assaad et al., 2021; Johansson et al., 2022; Shalit et al., 2017) • Worst-Case Guarantees (Jeong & Namkoong, 2020; Kern et al., 2024)	• Worst-Case Guarantees (Hatt et al., 2022, Kallus et al., 2022, Si et al., 2020, Si et al., 2023)
Label Bias	• Class-Conditional Label Noise (Natarajan et al., 2013; Yu et al., 2020) • Feature-Conditional Label Noise (Chen et al., 2021; Wang et al., 2021) • Multiple Proxy Variables (Boeschoten et al., 2021)	• Counterfactual Prediction under Measurement Error (Guerdan, Coston, Holstein, & Wu, 2023,Guerdan, Coston, Wu, & Holstein, 2023)	No dedicated methods specific to policy learning have been identified. Methods from other frameworks may be applicable.
Past Decision-Making	Unbiased estimation not possible if the outcome was influenced by interventions.	Requires causal identification. See Appendix B for an overview of ML- based CATE estimation methods.	Requires causal identification. See Appendix C for an overview of off-policy learning methods.
Competing Objectives & Constraints	• Scalar Utility Functions (Boutilier, 2013; Keeney & Raiffa, 1993) • Solutions along the Pareto Frontier (Hertweck et al., 2022; Pfisterer et al., 2019) and Model Multiplicity (Watson-Daniels et al., 2023) • Specific Model Constraints, such as Fairness	• Budget-Constrained Allocation (Ai et al., 2022; McFowland III et al., 2021; Tu et al., 2021) • Fairness Constraints (Kim & Zubizarreta, 2023; Mishler et al., 2021)	• Budget-Constrained Allocation (Huang & Xu, 2020; Sun, 2021; Xu et al., 2022) • Solutions along the Pareto Frontier (Rehill & Biddle, 2024) • Specific Model Constraints (Athey &

(continued on next page)

Table A1 (*continued*)

	Risk Prediction	Counterfactual Prediction	Off-Policy Learning
	(Hardt, Price, et al., 2016; Hort, Chen, Zhang, Harman and Sarro, 2023) and Inter- pretability (Molnar, 2022)		Wager, 2021), such as Fairness (Frauen et al., 2024)
Uncertainty Estimation for Human-in-the-Loop	• Model agnostic Uncertainty Estimation with Conformal Prediction (Angelopoulos & Bates, 2022; Papadopoulos, Proedrou, Vovk, & Gammerman, 2002b; Straitouri, Wang, Okati, & Rodriguez, 2023)	• Model agnostic Uncertainty Estimation for CATEs with weighted Conformal Prediction (Lei & Candès, 2021; Tibshirani et al., 2019) and conformal meta-learners (Alaa et al., 2023)	• Uncertainty in Policy Value (Wang et al., 2017) and Conformal Off-Policy Evaluation for Outcomes (Taufiq et al., 2022)

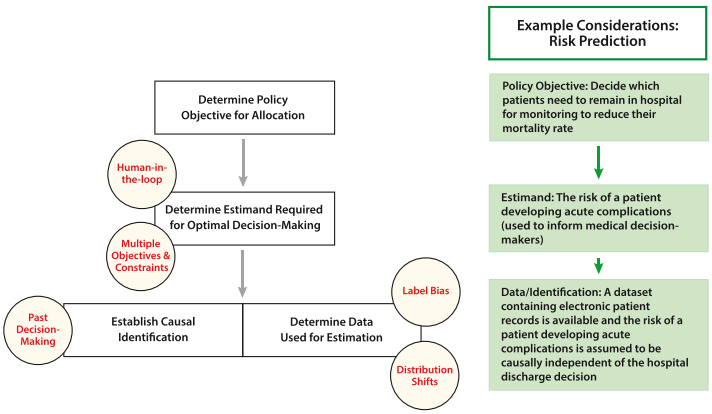


Fig. A1. Key Questions for Policy Makers in Selecting Risk Prediction as the Modeling Approach. Example inspired by algorithmic predictions of acute gastrointestinal bleeding (Alur et al., 2023).

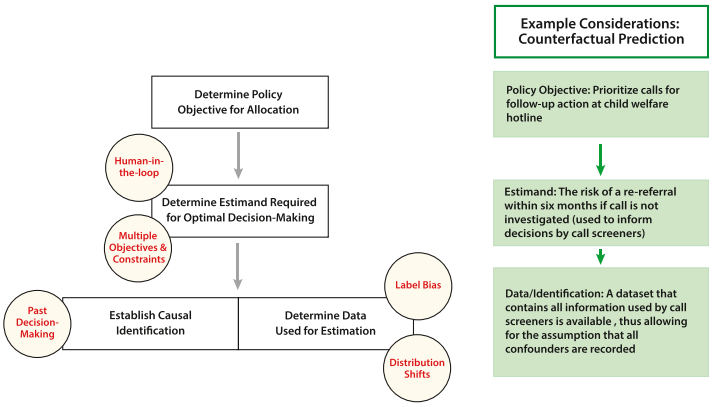


Fig. A2. Key Questions for Policy Makers in Selecting Counterfactual Prediction as the Modeling Approach. Example inspired by child abuse hotline screening in Allegheny County (Chouldechova et al., 2018; Coston et al., 2020).

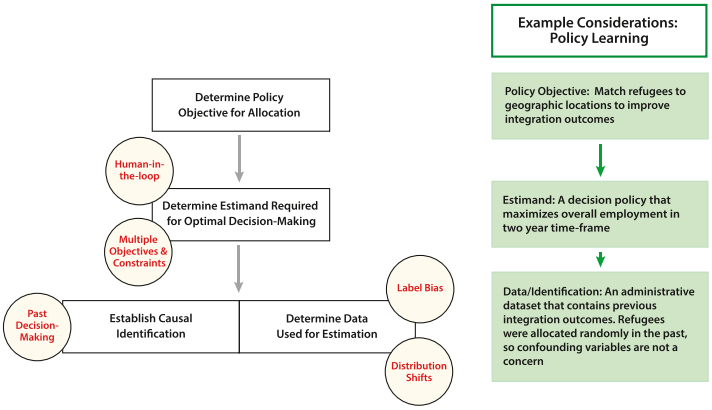


Fig. A3. Key Questions for Policy Makers in Selecting Policy Learning as the Modeling Approach. Example inspired by geographical matching of refugees to improve integration outcomes (Bansak et al., 2018).

Appendix B. CATE Estimation Methods

In recent years, several CATE estimation strategies have emerged, many of which employ nonparametric ML regression models to estimate the relationship between covariates, outcome and intervention (Caron et al., 2022; Lechner, 2023). ML models are generally well-suited for inferring complex non-linear relationships and handling a larger number of covariates, which can be vital for capturing heterogeneous effects. Model-agnostic meta-algorithms for CATE estimation enable the use of an arbitrary ML model as a base learner, such as random forests and neural networks (Caron et al., 2022; Curth & van der Schaar, 2021; Künzel et al., 2019).

One class of meta-learners initially aims to estimate both expected outcome functions $\mu_t(x)$, and computing the CATE as the difference between the estimates of these functions (Curth & van der Schaar, 2021). For example, S-learners treat the treatment indicator as an additional feature and estimate the potential outcomes with a single outcome regression $\mu(x,t) = \mathbb{E}[Y|X=x,T=t]$. T-learners use ML models to estimate $\mu_0(x) = \mathbb{E}[Y|X=x,T=0]$ and $\mu_1(x) = \mathbb{E}[Y|X=x,T=1]$ separately. However, both approaches can introduce significant bias, particularly when dealing with imbalanced treatment groups (Caron et al., 2022; Johansson et al., 2022; Künzel et al., 2019; Nie & Wager, 2020).

Alternative methods directly estimate the CATE function by first constructing pseudo-outcomes of the treatment effects (Caron et al., 2022; Curth & van der Schaar, 2021; Künzel et al., 2019). One prominent variant is the X-learner (Künzel et al., 2019), an extension of the T-learner, which can also be regarded as a special case of the RA-learner (Curth & van der Schaar, 2021). The Doubly-Robust learner (DR-learner) employs pseudo-outcomes constructed from both the conditional outcomes $\widehat{\mu}_t(x)$ and the propensity scores $\widehat{\pi}(x)$ (Kennedy, 2023). DR estimators have a long history in causal inference and missing data imputation (Bang & Robins, 2005; Funk et al., 2011; Kang & Schafer, 2007; Robins, Rotnitzky, & Zhao, 1994) and offer the advantage of consistency as long as either the propensity score model or conditional outcome models are correctly specified. Finally, the R-learner (Foster & Syrgkanis, 2023; Nie & Wager, 2020) involves the formulation a specific loss function after fitting several nuisance functions that can be separately minimized and regularized to estimate the CATE, drawing inspiration from the Robinson decomposition (Robinson, 1988). There also exists CATE estimation methods that adapt specific ML models (Jacob, 2021). For instance, causal forests, introduced by Athey, Tibshirani, and Wager (2019); Wager and Athey (2018), resemble the R-learner (Caron et al., 2022; Oprescu, Syrgkanis, & Wu, 2019; Post, van den Heuvel, Petkovic, & van den Heuvel, 2024).

A large body of literature analyzes the asymptotic and finite sample properties of different CATE learners (Caron et al., 2022; Curth & van der Schaar, 2021; Kennedy, 2023; Künzel et al., 2019; Salditt, Eckes, & Nestler, 2023). However, providing clear guidelines for determining the most suitable approach in real-world scenarios remains challenging. Which method will be most appropriate will generally depend on various factors, such as the level of confounding, the presence of high-dimensional covariates, the expected complexity of the CATE function compared to the individual outcomes and whether the treatment groups are strongly unbalanced. We recommend Curth and Van Der Schaar (2023) for a detailed discussion of the advantages and disadvantages of different CATE estimation strategies.

Appendix C. Off-Policy Learning

Common approaches to construct an estimator for the policy value from observational data involve using weighting techniques, where propensity scores are estimated to re-balance the data, making it resemble data generated under the target policy (Kallus, 2018; Swaminathan & Joachims, 2015). For instance, Kitagawa and Tetenov (2018) develop an algorithm that makes use of inverse propensity score weighting (IPW) to estimate $V(\pi)$ in a binary deterministic decision setting. Alternatively, some methods opt for direct estimation of the optimal treatment policy by fitting the outcome regression $\mathbb{E}[Y|X=x,T=t]$ and use the resulting estimates to optimize the policy value $\widehat{V}$ (Bennett & Kallus, 2020; Qian & Murphy, 2011). Doubly Robust (DR) methods combine the IPW and direct approach by using an augmented IPW (AIPW) loss (Robins et al., 1994). This requires estimating both the propensity scores and the outcome regression model (Athey & Wager, 2021; Dudík, Langford, & Li, 2011; Zhang, Tsiatis, Davidian, Zhang, & Laber, 2012). Several approaches have been proposed to relax the assumptions for causal identification, for example methods that address learning

policies under unmeasured confounding (Bennett & Kallus, 2019) or handle situations with limited overlap (Kallus, 2021).

References

Acharki, N., Lugo, R., Bertoncello, A., & Garnier, J. (2023). Comparison of Meta-learners for estimating multi-valued treatment heterogeneous effects. In *Proceedings of the 40th international conference on machine learning, ICML'23*. Honolulu, Hawaii, USA: JMLR.org.

Ai, M., Li, B., Gong, H., Yu, Q., Xue, S., Zhang, Y., … Jiang, P. (2022). Lbcf: A large-scale budget-constrained causal forest algorithm. In *Proceedings of the ACM Web Conference 2022, WWW '22, page 2310–2319*. New York, NY, USA: Association for Computing Machinery.

Alaa, A., Ahmad, Z., & van der Laan, M. (2023). *Conformal meta-learners for predictive inference of individual treatment effects. Advances in neural information processing systems* (vol. 36, pp. 47682–47703). Curran Associates, Inc.

Alexopoulos, C., Lachana, Z., Androutsopoulou, A., Diamantopoulou, V., Charalabidis, Y., & Loutsaris, M. A. (2019). How machine learning is changing E-government. In *In proceedings of the 12th international conference on theory and practice of electronic governance, pages 354–363*. New: York. Association for Computing Machinery.

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. *Frontiers in Big Data, 3*.

Alur, R., Laine, L., Li, D., Raghavan, M., Shah, D., & Shung, D. (2023). Auditing for human expertise. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *vol. 36. Advances in neural information processing systems* (pp. 79439–79468). Curran Associates, Inc.

Amarasinghe, K., Rodolfa, K. T., Lamba, H., & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy, 5*, Article e5.

Angelopoulos, A. N., & Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint*. arXiv:2107.07511.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., … Carin Duke, L. (2021). Counterfactual representation learning with balancing weights. In A. Banerjee, & K. Fukumizu (Eds.), *Proceedings of the 24th international conference on artificial intelligence and statistics, volume 130 of proceedings of machine learning research* (pp. 1972–1980). PMLR.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science (New York, N.Y.), 355*(6324), 483–485.

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Working paper 26463*. National Bureau of Economic Research.

Athey, S., Keleher, N., & Spiess, J. (2023). Machine learning who to nudge: Causal vs predictive targeting in a field experiment on student financial aid renewal. *arXiv preprint*. arXiv:2310.08672.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148–1178.

Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica, 89*(1), 133–161.

Bach, R. L., Kern, C., Mautner, H., & Kreuter, F. (2023). The impact of modeling decisions in statistical profiling. *Data & Policy, 5*, Article e32.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics. Journal of the International Biometric Society, 61*(4), 962–973.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science, 359*(6373), 325–329.

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics, 51*(2), 816–845.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data, 4*.

Bennett, A., & Kallus, N. (2019). Policy evaluation with latent confounders via optimal balance. In *, vol. 32. Advances in neural information processing systems*. Curran Associates, Inc.

Bennett, A., & Kallus, N. (2020). Efficient policy learning from surrogate-loss classification reductions. In H. D. III, & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning, volume 119 of proceedings of machine learning research* (pp. 788–798). PMLR.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., … Xiang, A. (2021). Uncertainty as a form of transparency: measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21* (pp. 401–413). New York, NY, USA: Association for Computing Machinery.

Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research, 10*(75), 2137–2155.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74*(5), 817–848.

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., & Ho, D. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 1479–1503). Seoul Republic of Korea: ACM.

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2018). *Cost-benefit analysis: Concepts and practice* (5 ed.). Cambridge University Press.

Boeschoten, L., van Kesteren, E.-J., Bagheri, A., & Oberski, D. L. (2021). Achieving fair inference using error-prone outcomes. *International Journal of Interactive Multimedia and Artificial Intelligence, 6*(5), 9–15.

Boutilier, C. (2013). Computational decision support regret-based models for optimization and preference elicitation. In T. R. Zentall, & P. H. Crowley (Eds.), *Comparative decision making* (pp. 423–453). Oxford University Press.

Caron, A., Baio, G., & Manolopoulou, I. (2022). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(3), 1115–1149.

Chen, J., Li, Z., & Mao, X. (2023). Learning under selective labels with data from heterogeneous decision-makers: An instrumental variable approach. *arXiv preprint*. arXiv:2306.07566.

Chen, P., Ye, J., Chen, G., Zhao, J., & Heng, P.-A. (2021). Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *, 35. Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 11442–11450).

Cheng, L., & Chouldechova, A. (2022). Heterogeneity in algorithm-assisted decision-making: A case study in child abuse hotline screening. *Proc. ACM Hum.-Comput. Interact., 6*(CSCW2).

Chiusi, F. (Ed.). (2020). *Automating society report 2020*. Germany: AlgorithmWatch gGmbH & Bertelsmann Stiftung. https://automatingsociety.algorithmwatch.org.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In S. A. Friedler, & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency, volume 81 of proceedings of machine learning research* (pp. 134–148). PMLR.

Chugunova, M., & Sele, D. (2023). Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *Academy of Management Proceedings, 2023*(1), 16472.

Cockx, B., Lechner, M., & Bollens, J. (2023). Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics, 80*, Article 102306.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '17* (pp. 797–806). New York, NY: USA. Association for Computing Machinery.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., … Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology, 8*(1), 4–36.

Coston, A., Kawakami, A., Zhu, H., Holstein, K., & Heidari, H. (2023). A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML), pages 690–704*.

Coston, A., Mishler, A., Kennedy, E. H., & Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on fairness, accountability, and transparency, FAT* '20* (pp. 582–593). New York, NY, USA: Association for Computing Machinery.

Coyle, D., & Weller, A. (2020). "Explaining" machine learning reveals policy challenges. *Science (New York, N.Y.), 368*(6498), 1433–1434.

Curth, A., & van der Schaar, M. (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th international conference on artificial intelligence and statistics* (pp. 1810–1818). PMLR.

Curth, A., & Van Der Schaar, M. (2023). In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International conference on machine learning* (pp. 6623–6642). PMLR.

Dai, J., & Brown, S. M. (2020). Label bias, label shift: fair machine learning with unreliable labels. In *, 12. NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*.

Das, I., & Dennis, J. E. (1997). A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural optimization, 14*(1), 63–69.

David, S. B., Lu, T., Luu, T., & Pal, D. (2010). Impossibility theorems for domain adaptation. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 129–136). JMLR Workshop and Conference Proceedings.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (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* (pp. 1–12). New York, NY, USA: Association for Computing Machinery.

Deb, K. (2011). Multi-objective optimisation using evolutionary algorithms: An introduction. In L. Wang, A. H. C. Ng, & K. Deb (Eds.), *Multi-objective evolutionary optimisation for product design and manufacturing* (pp. 3–34). London: Springer.

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization, EAAMO '23*. New York, NY, USA: Association for Computing Machinery.

4 Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

Desiere, S., & Struyven, L. (2021). Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy, 50*(2), 367–385.

Dickerman, B. A., & Hernán, M. A. (2020). Counterfactual prediction is not only for causal inference. *European Journal of Epidemiology, 35*(7), 615–617.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint* (arXiv:1702.08608).

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), Article eaao5580.

Duchi, J. C., & Namkoong, H. (2021). Learning models with uniform performance via Distributionally robust optimization. *The Annals of Statistics, 49*(3), 1378–1406.

Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th international conference on international conference on machine learning, ICML'11* (pp. 1097–1104). Madison, WI, USA: Omnipress.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., … Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, Article 101994.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science, 32*(2), 249–264.

Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize". *Management Science, 68*(1), 9–26.

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop – Legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law, 31*(1), 123–153.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by algorithm: Artificial intelligence in federal administrative agencies* (pp. 20–54). NYU School of Law, Public Law Research Paper.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 160–171). PMLR.

EU Commission. (2021). *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Ec, Com* (p. 206).

Fang, T., Lu, N., Niu, G., & Sugiyama, M. (2020). Rethinking importance weighting for deep learning under distribution shift. In *, 33. Advances in neural information processing systems* (pp. 11996–12007). Curran Associates, Inc.

Fernández-Loría, C., & Loría, J. (2023). Causal scoring: A framework for effect estimation, effect ordering, and effect classification. *arXiv preprint*. arXiv:2206.12532.

Fernández-Loría, C., & Provost, F. (2022a). Causal decision making and causal effect estimation are not the same… And why it matters. *Informs Journal on Data science, 1*(1), 4–16.

Fernández-Loría, C., & Provost, F. (2022b). Rejoinder to "causal decision making and causal effect estimation are not the same… And why it matters". *Informs journal on data science, 1*(1), 23–26.

Fogliato, R., Chouldechova, A., & G'Sell, M. (2020). Fairness evaluation in presence of biased Noisy labels. In *Proceedings of the twenty third international conference on artificial intelligence and statistics* (pp. 2325–2336). PMLR.

Foster, D. J., & Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics, 51*(3), 879–908.

Fountain, J. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly, 39*(2), Article 101645.

Frauen, D., Melnychuk, V., & Feuerriegel, S. (2024). *Fair off-policy learning from observational data, 235*, 13943–13972.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*(7), 761–767.

Gerdon, F., Bach, R. L., Kern, C., & Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society, 9*(1), 20539517221089305.

Gibbs, I., & Candes, E. (2021). Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *34. Advances in neural information processing systems* (pp. 1660–1672). Curran Associates, Inc.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association: JAMIA, 19*(1), 121–127.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., & Kauermann, G. (2023). Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint*. arXiv:2305.16703.

Guerdan, L., Coston, A., Holstein, K., & Wu, Z. S. (2023). Counterfactual prediction under outcome measurement error. In *2023 ACM conference on fairness, accountability, and transparency* (pp. 1584–1598). Chicago IL USA: ACM.

Guerdan, L., Coston, A., Wu, Z. S., & Holstein, K. (2023). Ground(Less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT '23* (pp. 688–704). New York, NY, USA: Association for Computing Machinery.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic Classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science, ITCS '16* (pp. 111–122). New York, NY, USA: Association for Computing Machinery.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *29. Advances in neural information processing systems*. Curran Associates, Inc.

Hatamyar, J., & Kreif, N. (2023). Policy learning with rare outcomes. *arXiv preprint*. arXiv:2302.05260.

Hatt, T., Tschernutter, D., & Feuerriegel, S. (2022). Generalizing off-policy learning under sample selection bias. In *Proceedings of the thirty-eighth conference on uncertainty in artificial intelligence* (pp. 769–779). PMLR.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., … Roijers, D. M. (2022). A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems, 36*(1), 26.

Hebert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th international conference on machine learning* (pp. 1939–1948). PMLR.

Hedegaard, L., Sheikh-Omar, O. A., & Iosifidis, A. (2021). Supervised domain adaptation: A graph embedding perspective and a rectified experimental protocol. *IEEE Transactions on Image Processing, 30*, 8619–8631.

Hertweck, C., Baumann, J., Loi, M., Viganò, E., & Heitz, C. (2022). A justice-based framework for the analysis of algorithmic fairness-utility trade-offs. *arXiv preprint*. arXiv:2206.02891.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2023). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal of Responsible Computing, 1*(2), Article 11.

Huang, X., & Xu, J. (2020). Estimating individualized treatment rules with risk constraint. *Biometrics, 76*(4), 1310–1318.

Hüllermeier, E. (2021). Prescriptive machine learning for automated decision making: Challenges and opportunities. *arXiv preprint*. arXiv:2112.08268.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*(3), 457–506.

Hwang, K. (2016). Cost-benefit analysis: Its usage and critiques. *Journal of Public Affairs, 16*(1), 75–80.

Jacob, D. (2021). CATE meets ML – The conditional average treatment effect and machine learning. *Digital Finance, 3*(2), 99–148.

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*(3), 371–377 (Open and Smart Governments: Strategies, Tools, and Experiences).

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research, 70*, 338–345.

Jeong, S., & Namkoong, H. (2020). Robust causal inference under covariate shift via worst-case subpopulation treatment effects. In *Proceedings of thirty third conference on learning theory* (pp. 2079–2084). PMLR.

Johansson, F. D., Shalit, U., Kallus, N., & Sontag, D. (2022). Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research, 23*(166), 1–50.

Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. In *, 48. Proceedings of the 33rd international conference on international conference on machine learning* (pp. 3020–3029). New York, NY, USA: JMLR.org. ICML'16.

Kaiser, P., Kern, C., & Rügamer, D. (2022). Uncertainty-aware predictive modeling for fair data-driven decisions. *arXiv preprint*. arXiv:2211.02730.

Kallus, N. (2018). Balanced policy evaluation and learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *31. Advances in neural information processing systems*. Curran Associates, Inc.

Kallus, N. (2021). More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association, 116*(534), 646–658.

Kallus, N., Mao, X., Wang, K., & Zhou, Z. (2022). Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research* (pp. 10598–10632). PMLR.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4).

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.

Kennedy, E. H. (2020). Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics, 16*(1).

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics, 17*(2), 3008–3049.

Kern, C., Kim, M., & Zhou, A. (2024). *Multi-cate: Multi-accurate conditional average treatment effect estimation robust to unknown covariate shifts*.

Kerrigan, D., Hullman, J., & Bertini, E. (2021). A survey of domain knowledge elicitation in applied machine learning. *Multimodal Technologies and Interaction, 5*(12), 73.

Kim, K., & Zubizarreta, J. R. (2023). Fair and robust estimation of heterogeneous treatment effects for policy learning. In *International conference on machine learning* (pp. 16997–17014). PMLR.

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, AIES '19* (pp. 247–254). New York, NY, USA: Association for Computing Machinery.

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., & Reingold, O. (2022). Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences, 119*(4), Article e2108097119.
Kim, M. P., & Perdomo, J. C. (2023). *Making decisions under outcome performativity*. In 14th Innovations in. In *, 251. Theoretical Computer Science Conference. Leibniz International Proceedings in Informatics (LIPIcs)* (pp. 1–79:15,). Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
Kitagawa, T., & Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica, 86*(2), 591–616.
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions*. *The Quarterly Journal of Economics, 133*(1), 237–293.
Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–495.
Kolkman, D. (2020). The usefulness of algorithmic models in policy making. *Government Information Quarterly, 37*(3), Article 101488.
Körtner, J., & Bonoli, G. (2023). Predictive algorithms in the delivery of public employment services. In *Handbook of labour market policy in advanced democracies* (pp. 387–398). Edward Elgar Publishing.
Kouw, W. M., & Loog, M. (2018). An introduction to domain adaptation and transfer learning. *arXiv preprint*. arXiv:1812.11806.
Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research, 297*(3), 1083–1094.
Kraus, M., Feuerriegel, S., & Saar-Tsechansky, M. (2024). Data-driven allocation of preventive care with application to diabetes mellitus type ii. *Manufacturing & Service Operations Management, 26*(1), 137–153.
Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences, 116*(10), 4156–4165.
Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2022). From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology, 7*, Article 883999.
Kuzmanovic, M., Frauen, D., Hatt, T., & Feuerriegel, S. (2024). Causal machine learning for cost-effective allocation of development aid. *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 5283–5294). Association for Computing Machinery.
Kuzmanovic, M., Hatt, T., & Feuerriegel, S. (2023). Estimating conditional average treatment effects with missing treatment information. In F. Ruiz, J. Dy, & J.-W. van de Meent (Eds.), *Proceedings of the 26th international conference on artificial intelligence and statistics, volume 206 of proceedings of machine learning research* (pp. 746–766). PMLR.
Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading Black box explanations. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society, AIES '20* (pp. 79–85). New York, NY, USA: Association for Computing Machinery.
Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 275–284). Halifax NS Canada: ACM.
Laux, J., Wachter, S., & Mittelstadt, B. (2023). *Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk*. Regulation & Governance.
Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals Deal with opacity when using AI for medical diagnosis. *Organization Science, 33*(1), 126–148.
Lechner, M. (2023). Causal machine learning and its use for public policy. *Swiss Journal of Economics and Statistics, 159*(1), 8.
Lei, L., & Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 83*(5), 911–938.
Lenert, M. C., Matheny, M. E., & Walsh, C. G. (2019). Prognostic models will be victims of their own success, unless. *Journal of the American Medical Informatics Association: JAMIA, 26*(12), 1645–1650.
Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science, 17*(1), 309–334.
Lin, L., Sperrin, M., Jenkins, D. A., Martin, G. P., & Peek, N. (2021). A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnostic and Prognostic Research, 5*(1), 3.
Lipton, Z., Wang, Y.-X., & Smola, A. (2018). Detecting and correcting for label shift with Black box predictors. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning, volume 80 of proceedings of machine learning research* (pp. 3122–3130). PMLR.
Liu, A., & Ziebart, B. (2014). Robust classification under sample selection Bias. In *, vol. 27. Advances in neural information processing systems*. Curran Associates, Inc.
Luedtke, A. R., & van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics, 44*(2), 713–742.
Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19.
Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica, 72*(4), 1221–1246.
Matzat, L. (Ed.). (2019). *Atlas of automation –Automated decision-making and participation in Germany* (1st edition). AW AlgorithmWatch gGmbH. https://atlas.algorithmwatch.org/en/.
Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended consequences of introducing ai systems for decision making. *MIS Quarterly Executive, 19*(4).
McFowland, E., III, Gangarapu, S., Bapna, R., & Sun, T. (2021). A prescriptive analytics framework for optimal policy deployment using heterogeneous treatment effects. *MIS Quarterly, 45*(4).
McKay, C. (2020). Predicting risk in criminal procedure: Actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice, 32*(1), 22–39.
Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly, 81*(S1), 250–271.
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
Mishler, A., Kennedy, E. H., & Chouldechova, A. (2021). Fairness in risk assessment instruments: post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 386–400).
Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*(1), 141–163.
Mitrou, L., Janssen, M., & Loukis, E. (2022). Human control and discretion in ai-driven decision-making in government. In *Proceedings of the 14th international conference on theory and practice of electronic governance, ICEGOV '21* (pp. 10–16). New York, NY, USA: Association for Computing Machinery.
Molnar, C. (2022). *Interpretable machine learning: A guide for making Black box models explainable* (2 ed.) https://christophm.github.io/interpretable-ml-book/.
Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition, 45*(1), 521–530.
Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080.
Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with Noisy labels. In *, 26. Advances in neural information processing systems*. Curran Associates, Inc.
Nie, X., & Wager, S. (2020). Quasi-Oracle estimation of heterogeneous treatment effects. *Biometrika, 108*(2), 299–319.
van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, 101714.
Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*, 97–105.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.), 366*(6464), 447–453.
Oprescu, M., Syrgkanis, V., & Wu, Z. S. (2019). Orthogonal random Forest for causal inference. In *Proceedings of the 36th international conference on machine learning* (pp. 4932–4941). PMLR.
Pagan, N., Baumann, J., Elokda, E., De Pasquale, G., Bolognani, S., & Hannák, A. (2023). A classification of feedback loops and their relation to biases in automated decision-making systems, *7. Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization* (pp. 14–pages). Association for Computing Machinery.
Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002a). Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning, ECML '02* (pp. 345–356). Berlin, Heidelberg: Springer-Verlag.
Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002b). Inductive confidence machines for regression. In *Proceedings of the 13th European conference on machine learning, ECML'02* (pp. 345–356). Berlin: Heidelberg. Springer-Verlag.
Papalexopoulos, T. P., Bertsimas, D., Cohen, I. G., Goff, R. R., Stewart, D. E., & Trichakis, N. (2022). Ethics-by-design: Efficient, fair and inclusive resource allocation using machine learning. *Journal of Law and the Biosciences, 9*(1), lsac012.
Passi, S., & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 39–48). Atlanta GA USA: ACM.
Peeters, R., & Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture–the case of the dutch civil registry's master data management system. *Government Information Quarterly, 35*(2), 175–183.
Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big data and AI – A transformational shift for government: So, what next for research? *Public Policy and Administration, 35*(1), 24–44.
Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In H. D. III, & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning, volume 119 of proceedings of machine learning research* (pp. 7599–7609). PMLR.
Perdomo, J. C. (2024). *The relative value of prediction in algorithmic decision making, 235*, 40439–40460.
Perdomo, J. C., Britton, T., Hardt, M., & Abebe, R. (2023). Difficult lessons on social prediction from Wisconsin public schools. *arXiv preprint*. arXiv:2304.06205.
Pfisterer, F., Coors, S., Thomas, J., & Bischl, B. (2019). Multi-objective automatic machine learning with autoxgboostmc. *arXiv preprint*. arXiv:1908.10796.
Post, R., van den Heuvel, I., Petkovic, M., & van den Heuvel, E. (2024). Flexible machine learning estimation of conditional average treatment effects: A blessing and a curse. *Epidemiology, 35*(1), 32–40.
Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgenson, E., Mansour, R., & Ghani, R. (2015). Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. In *Proceedings of the 21th ACM*

4 Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector

SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2039–2047). Sydney NSW Australia: ACM.
Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems, CHI '21* (pp. 1–52). New York, NY, USA: Association for Computing Machinery.
Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., … Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence, 2*(7), 369–375.
Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics, 39*(2), 1180–1210.
Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). *Dataset shift in machine learning*. MIT Press.
Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). Covariate shift by kernel mean matching. In *Dataset shift in machine learning* (pp. 131–160). MIT Press.
Rambachan, A., Coston, A., & Kennedy, E. (2022). Counterfactual risk assessments under unmeasured confounding. *arXiv preprint*. arXiv:2212.09844.
Rehill, P., & Biddle, N. (2024). Policy learning for many outcomes of interest: Combining optimal policy trees with multi-objective bayesian optimisation. *Comput Econ*.
Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some Regressors are not always observed. *Journal of the American Statistical Association, 89*(427), 846–866.
Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica, 56* (4), 931–954.
Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research, 48*, 67–113.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.
Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys, 16*, 1–85. none.
Salditt, M., Eckes, T., & Nestler, S. (2023). A tutorial introduction to heterogeneous treatment effect estimation with meta-learners. *Administration and Policy in Mental Health and Mental Health Services Research*, 1–24.
Schulam, P., & Saria, S. (2017). Reliable decision support using counterfactual models. In *, 30. Advances in neural information processing systems*. Curran Associates, Inc.
Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly, 85* (S1), 399–422.
Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation Bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys, 55*(13s), 1–39.
Shalit, U. (2022). Commentary on "causal decision making and causal effect estimation are not the same… And why it matters". *Informs journal on data Science, 1*(1), 19–20.
Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th international conference on machine learning* (pp. 3076–3085). PMLR.
Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference, 90* (2), 227–244.
Si, N., Zhang, F., Zhou, Z., & Blanchet, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *Proceedings of the 37th international conference on machine learning* (pp. 8884–8894). PMLR.
Si, N., Zhang, F., Zhou, Z., & Blanchet, J. (2023). Distributionally robust batch contextual bandits. *Management Science, 69*(10), 5772–5793.
Singh, K., Valley, T. S., Tang, S., Li, B. Y., Kamran, F., Sjoding, M. W., … Nallamothu, B. K. (2021). Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Annals of the American Thoracic Society, 18*(7), 1129–1137.
Straitouri, E., & Rodriguez, M. G. (2024). Designing decision support systems using counterfactual prediction sets. In *, 235. Proceedings of the 41st international conference on machine learning, in Proceedings of machine learning research* (pp. 46722–46744).
Straitouri, E., Wang, L., Okati, N., & Rodriguez, M. G. (2023). Improving expert predictions with conformal prediction. In *Proceedings of the 40th international conference on machine learning, volume 202 of ICML'23* (pp. 32633–32653). Honolulu, Hawaii, USA: JMLR.org.
Subbaswamy, A., Chen, B., & Saria, S. (2022). A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference, 10*(1), 64–89.
Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., & Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *, Vol. 20. Advances in neural information processing systems*. Curran Associates, Inc.
Sullivan, T. (2015). *Introduction to uncertainty quantification, volume 63 of Texts in Applied Mathematics*. Cham: Springer International Publishing.
Sun, H., Munro, E., Kalashnov, G., Du, S., & Wager, S. (2024). Treatment allocation under uncertain costs. *arXiv preprint*. arXiv:2103.11066.
Sun, L. (2021). Empirical welfare maximization with constraints. *arXiv preprint, 2*. arXiv: 2103.15298.
Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 36* (2), 368–383.
Swaminathan, A., & Joachims, T. (2015). Counterfactual risk minimization. In *Proceedings of the 24th international conference on world wide web* (p. 939). Florence Italy: ACM.
Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., & Doucet, A. (2022). Conformal off-policy prediction in contextual bandits. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Vol. 35. Advances in neural information processing systems* (pp. 31512–31524). Curran Associates, Inc.
Tibshirani, R. J., Foygel Barber, R., Candes, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Vol. 32. Advances in neural information processing systems*. Curran Associates, Inc.
Tourangeau, R. (2019). Surveying hard-to-survey populations despite the unfavorable environment. *American Journal of Public Health, 109*(10), 1326.
Tu, Y., Basu, K., DiCiccio, C., Bansal, R., Nandy, P., Jaikumar, P., & Chatterjee, S. (2021). Personalized treatment selection using causal heterogeneity. In *Proceedings of the Web Conference 2021, WWW '21* (pp. 1574–1585). New York, NY, USA: Association for Computing Machinery.
Vaithianathan, R., Benavides-Prado, D., Dalton, E., Chouldechova, A., & Putnam-Hornstein, E. (2021). Using a machine learning tool to support high-stakes decisions in child protection. *AI Magazine, 42*(1), 53–60.
Vaithianathan, R., Kulick, E., Putnam-Hornstein, E., & Benavides-Prado, D. (2019). Allegheny family screening tool: Methodology, version 2. *Center for Social Data Analytics*, 1–22.
Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research, 40*(1), 105–137.
Van Geloven, N., Swanson, S. A., Ramspek, C. L., Luijken, K., Van Diepen, M., Morris, T. P., … Le Cessie, S. (2020). Prediction meets causal inference: The role of treatment in clinical prediction models. *European Journal of Epidemiology, 35*(7), 619–630.
Vegetabile, B. G. (2021). On the distinction between "conditional average treatment effects" (CATE) and "individual treatment effects" (ITE) under Ignorability assumptions. *arXiv preprint*. arXiv:2108.04939.
Vodrahalli, K., Gerstenberg, T., & Zou, J. Y. (2022). Uncalibrated models can improve human-AI collaboration. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Vol. 35. Advances in neural information processing systems* (pp. 4004–4016). Curran Associates, Inc.
Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic learning in a random world*. Cham: Springer International Publishing.
Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.
Wang, A., Kapoor, S., Barocas, S., & Narayanan, A. (2023). Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT '23* (p. 626). New York, NY, USA: Association for Computing Machinery.
Wang, J., Liu, Y., & Levy, C. (2021). Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 526–536).
Wang, Y.-X., Agarwal, A., & Dudík, M. (2017). Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th international conference on machine learning - volume 70, ICML'17* (pp. 3589–3597). JMLR.org.
Watson-Daniels, J., Barocas, S., Hofman, J. M., & Chouldechova, A. (2023). Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT '23* (pp. 297–311). New York, NY, USA: Association for Computing Machinery.
Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery, 30*(4), 964–994.
Wen, J., Yu, C.-N., & Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proceedings of the 31st international conference on machine learning* (pp. 631–639). PMLR.
Wirick, D. (2011). *Public-sector project management: Meeting the challenges and achieving results*. John Wiley & Sons.
Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration, 42* (7), 596–615.
Xu, Y., Greene, T. H., Bress, A. P., Sauer, B. C., Bellows, B. K., Zhang, Y., … Shen, J. (2022). Estimating the optimal individualized treatment rule from a cost-effectiveness perspective. *Biometrics, 78*(1), 337–351.
Yang, S., & Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science, 3*, 625–650.
Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making, 32*(4), 403–414.
Yu, X., Liu, T., Gong, M., Zhang, K., Batmanghelich, K., & Tao, D. (2020). Label-noise robust domain adaptation. In *Proceedings of the 37th international conference on machine learning* (pp. 10913–10924). PMLR.
Yu, Y.-L., & Szepesvári, C. (2012). Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th international Coference on international conference on machine learning, ICML'12* (pp. 1147–1154). Madison, WI, USA: Omnipress.
Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th international conference on artificial intelligence and statistics* (pp. 962–970). PMLR.
Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., & Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat, 1*(1), 103–114.
Zhang, Y., Shi, C., & Luo, S. (2023). Conformal off-policy prediction. In F. Ruiz, J. Dy, & J.-W. van de Meent (Eds.), *Proceedings of the 26th international conference on artificial intelligence and statistics, volume 206 of proceedings of machine learning research* (pp. 2751–2768). PMLR.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.
Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly, 38*(3), Article 101577.
Zhang, J., Menon, A., Veit, A., Bhojanapalli, S., Kumar, S., & Sra, S. (2021). Coping with label shift via distributionally robust optimisation. In *In Proceedings of the 9th international conference on learning representations (ICLR 2021)*.

Unai Fischer-Abaigar is a PhD Student at the Department of Statistics at LMU Munich, Munich Machine Learning Center and the Konrad Zuse School of Excellence in Reliable AI. His research leverages machine learning to enhance the reliability of high-stakes decision-making.

Christoph Kern is Junior Professor of Social Data Science and Statistical Learning at LMU Munich, Research Assistant Professor at the Joint Program in Survey Methodology (JPSM) at the University of Maryland and Project Director at the Mannheim Centre for European Social Research (MZES). His research focuses on the social impacts of algorithmic decision-making and on methodol- ogy to mitigate algorithmic unfairness and improve training data quality.

Noam Barda is a physician, with a specialty in public health and epidemiology. He is an Associate Professor at the Department of Software and Information Systems Engineering and the Department of Epidemiology, Biostatistics and Community Health Sciences at Ben-Gurion University of the Negev.

Frauke Kreuter is the Professor of Statistics and Data Science in Social Sciences and the Humanities at the Ludwig-Maximilians-University of Munich, Germany; Co-director of the Social Data Science Center (SoDa), and faculty member in the Joint Program in Survey Methodology (JPSM) at the University of Maryland, USA; and until recently head of the statistical methods group at the Institute for Employment Research (IAB) in Nuremberg, Germany.

5 Algorithms for reliable decision-making need causal reasoning

Publication

C. Kern, U. Fischer-Abaigar, J. Schweisthal, D. Frauen, R. Ghani, S. Feuerriegel, M. van der Schaar, and F. Kreuter (2025). "Algorithms for Reliable Decision-Making Need Causal Reasoning". *Nature Computational Science* 5:5, pp. 356–360. ISSN: 2662-8457. DOI: 10.1038/s43588-025-00814-9 Reproduced with permission from Springer Nature.

Author Contributions CK led the project as first author and was responsible for correspondence with the editor. CK, UFA, and SF initiated and conceptualized the paper and wrote the original draft. UFA conceptualized and created the figures and wrote the initial versions of Sections 1 and 2 and the subsection on performativity in Section 3. JS and DF contributed to the original draft of Section 3. RG, MvdS, and FK contributed to reviewing and editing. All authors contributed to revisions

Comment

https://doi.org/10.1038/s43588-025-00814-9

Algorithms for reliable decision-making need causal reasoning

Christoph Kern, Unai Fischer-Abaigar, Jonas Schweisthal, Dennis Frauen, Rayid Ghani, Stefan Feuerriegel, Mihaela van der Schaar & Frauke Kreuter

Check for updates

Decision-making inherently involves cause–effect relationships that introduce causal challenges. We argue that reliable algorithms for decision-making need to build upon causal reasoning. Addressing these causal challenges requires explicit assumptions about the underlying causal structure to ensure identifiability and estimatability, which means that the computational methods must successfully align with decision-making objectives in real-world tasks.

Algorithmic decision-making (ADM) has become common in a wide range of domains, including precision medicine, manufacturing, education, hiring, the public sector, and smart cities. At the core of ADM systems are data-driven models that learn from data to recommend decisions, often with the goal of maximizing a defined utility function[1]. For example, in smart city contexts, ADM is frequently used to optimize traffic flow through predictive models that analyze real-time data, thereby reducing congestion and improving urban mobility. Another prominent application area for ADM are normative decision support systems (often subsumed under 'prescriptive analytics') or, more recently, artificial intelligence (AI) agents that either inform or automatically execute managerial and operational decisions in industry. Yet, the applications of ADM to high-stakes decisions face safety and reliability issues[1–3]. Often, the objectives of ADM systems fail to align with the nuanced goals of real-world decision-making, thus creating a tension between the potential of ADM and the risk of harm and failure. Especially when deployed in dynamic, real-world environments, ADM can amplify systemic disadvantages for vulnerable communities and lead to flawed decisions.

In this Comment, we argue that reliable algorithmic decision-making – systems that perform safely and robustly under deployment conditions – must be grounded in causal reasoning.

Why ADM is rooted in causal reasoning

Decisions involve cause–effect relationships. Making the right decision – whether selecting the best treatment for a patient, choosing the most effective marketing strategy, or optimizing traffic signals in a smart city – requires understanding how a decision will influence the outcome of interest. Unlike a passive observation, the outcome is often directly shaped by the decision itself. For example, choosing a treatment to reduce a patient's mortality risk or adjusting traffic signals to minimize congestion directly affects the outcome being optimized. This distinguishes decision problems from pure associations, since decisions present interventions that eventually produce effects.

A fundamental challenge in modeling decision-making is that we can only observe the outcome for the decision that was actually taken – not what would have happened under alternative interventions (the fundamental problem of causal inference[4]). The latter is unobservable and thus referred to as the counterfactual outcome. Estimating and comparing the effects of different decisions requires reasoning about these counterfactuals. For example, if a patient receives a treatment and recovers, we cannot observe what their outcome would have been without the treatment. As a remedy, causal assumptions are needed that connect the observed data and the cause–effect relationships to express and eventually estimate counterfactual outcomes[5].

Causal assumptions are needed for modeling decision-making. Starting with clear assumptions about the data-generating process and causal relationships is crucial for establishing valid causal estimates. Such assumptions allow one to link a causal estimand (the quantity we aim to estimate to guide decision-making) to a statistical estimand (a quantity that can, in principle, be estimated from observed data). This distinction is critical: the causal estimand reflects our real objective, while the statistical estimand serves as the counterpart that we aim to estimate with our model. Without clearly defining the causal estimand first, it becomes impossible to ensure that our ADM model will align with the objective in our real-world tasks. Moreover, if the statistical estimand does not match the causal estimand, we risk producing unreliable estimates, leading to suboptimal or even incorrect decisions.

The above distinction between causal and statistical estimands highlights two key challenges: identifiability and estimatability. Identifiability refers to whether the causal estimand – the true effect of a decision – can be expressed as a function of observable data under the given assumptions[6]. Without identifiability, the effects of decisions, and therefore the corresponding outcomes, remain ambiguous, no matter how much data is available. In public policy, for example, large-scale administrative data might be available for ADM model training but these data rarely allow to recover the nuanced decision-rules based on which caseworkers have administered interventions in the past. This ambiguity is particularly problematic since it undermines the reliability of ADM models in practice. Without identifiability, any ADM model aiming to learn the correct effect of the decision remains biased. As a result, ADM models may produce incorrect or suboptimal decisions. In medicine, for example, both health outcomes and decisions for (costly) treatments may depend on unobserved factors such as patients' socio-economic status, preventing the ability to learn the true effect of treatment from the observed relationships. Estimatability, on the other hand, deals with whether the statistical estimand – once identified – can be reliably computed from finite data. Even when

a causal estimand is identifiable, estimating it is usually a statistically challenging task. Practical issues such as limited sample size, measurement noise, and model misspecification can hinder the ability to produce precise and unbiased estimates. Importantly, identifiability is a prerequisite for estimatability: without identifiability, the task of estimation becomes meaningless, as there is no assurance that the statistical estimand corresponds to the actual effect of a decision, meaning that an ADM model may not optimize against the utility of interest.

Generally, different frameworks are used to formalize causal assumptions[7], with two prominent approaches being structural causal models (SCMs) and the potential outcomes framework. While their methods and notation for specifying and identifying causal effects differ, they are often regarded as essentially equivalent in their core principles. SCMs use tools like causal graphs together with functional causal mechanisms to explicitly represent the cause–effect relationships, while, in contrast, the potential outcomes framework focuses on defining the outcomes that would occur under alternative decisions — capturing what would happen under each choice. To enable valid causal reasoning, several key assumptions about cause–effect relationships are often made[4]. For example, unconfoundedness assumes there are no hidden confounders that simultaneously influence both the decision and the outcome. Confounders in medical treatment decisions, for example, can include disease severity, patient characteristics, and access to healthcare. Positivity ensures that every decision has a nonzero probability of being observed for all individuals, providing sufficient data for comparison. Additionally, the stable unit treatment value assumption (SUTVA) essentially posits that an individual's outcome is influenced only by their own decision and not by the decisions of others.

Towards causal decision-making

Utilizing causal reasoning for reliable decision-making. To ensure reliable decisions with ADM, developers must first evaluate whether identifiability holds — that is, whether the causal estimand can be expressed using observable data and that the statistical estimand corresponds to the causal estimand. This requires clearly stating and justifying the assumptions about the data-generating process and causal relationships. By taking this step, developers can help align the objectives of ADM systems with the real-world goals they are meant to support. While this alone cannot guarantee safe decision-making, it is a critical prerequisite for ensuring that decision-making is both effective and reliable (Fig. 1).

An important step is to carefully evaluate whether the necessary causal assumptions hold in practice. For instance, when developers of ADM systems have access to randomized data from experiments, unconfoundedness is often ensured by design, making causal effects identifiable. For example, in a smart city context, randomized control trials might be performed to test the impact of adaptive traffic light systems on reducing congestion. In some cases, data from natural experiments may be available, for example, due to temporal malfunctions in operational processes or regional variation in the roll-out of policy programs. However, access to randomized data is rare, particularly in high-stakes settings where randomization is impractical or unethical. Instead, many ADM systems from practice rely on observational data generated by complex processes where interventions were assigned based on historical policies (for instance, expert judgment, organizational guidelines, and institutional policies). In these cases, guaranteeing causal assumptions can be challenging and often requires input from domain experts and additional contextual knowledge.

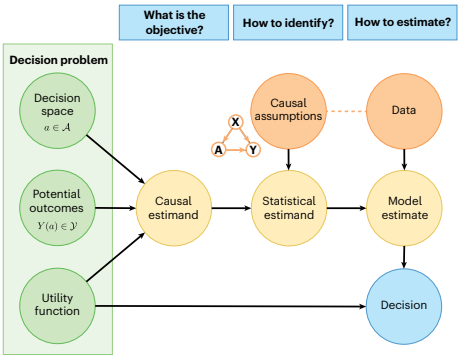


Fig. 1 | Automated decision-making requires causal assumptions to ensure reliable decisions. This figure shows the key components needed to set up an automated decision-making system. Arrows indicate how one component informs or depends on the next. Green nodes describe the decision problem: what the goal is, what decisions can be made, and what outcomes are affected. Yellow nodes represent the quantities used to connect this problem to data: the causal estimand (the effect we want to know), the statistical estimand (what we can estimate from data), and the model estimate (what we actually compute). Orange nodes represent external information needed to make this link: data and causal assumptions (such as no unmeasured confounders of decision A and outcome Y given covariates X) made about the data (dashed line). In practice, this framework can be extended to incorporate considerations such as operational constraints (for instance, resource constraints) or fairness constraints.

It is important to emphasize that ignoring causal assumptions simply because they are difficult to guarantee is not a solution. Failing to make causal assumptions explicit does not mean they do not exist — on the contrary, neglecting them can lead to flawed and unreliable decision-making. For example, naively training models on observational data generated by historical decision-making processes can result in systematically underestimating true risks[8]. Even ADM models that do not explicitly target counterfactual outcomes — or that do not state the causal assumptions — often rely on such causal assumptions implicitly. A prominent example is prediction-based decision rules, where a predictive model is trained and decisions are made based on a specified threshold[9]. Here, the causal assumptions are rarely stated but such ADM models typically assume independence between the outcome and decision or rely on a known model of treatment effects given the predicted baseline outcome. A common example is hospital discharge decisions, where the decision to discharge or retain a patient typically does not influence the underlying disease progression. In such contexts, predictive associations alone may suffice for making effective decisions. Nonetheless, explicitly clarifying these assumptions is crucial, as it ensures that the decision-makers are aware of potential limitations or scenarios where associations alone might fail due to unnoticed feedback loops.

We now offer a few illustrative examples. In prediction-based decision rules, the goal is often to intervene in cases with a high predicted risk of adverse outcomes if no help were provided. For instance, in

healthcare, a risk prediction model may trigger an alarm for patients where their health status is likely to deteriorate without treatment, thus allowing for targeted interventions. Here, the prediction – and thus the decision rule – is learned from historical data under interventions, and, hence, identification typically relies on the assumption of unconfoundedness or that there is sufficient variation in the historical decision-making. In offline policy learning, the objective is to maximize expected outcomes under a given decision policy, which maps observed covariates to actions. For example, the task might involve determining optimal traffic signal timings at intersections to minimize average travel time across the city. Achieving this goal requires assuming, for instance, unconfoundedness across all decisions so that intervention effects can be reliably estimated and the policy can optimize the expected utility. In reinforcement learning, the focus extends to optimizing the expected return over a time horizon, usually in an online setting. For example, adaptive traffic control systems might iteratively adjust signal timings based on real-time data to minimize congestion throughout the day. Here, identification is often naturally satisfied because decisions (for instance, signal changes) are assigned sequentially by a known policy. However, challenges arise in partially observed settings, such as unpredictable events, which require further effort to come up with an identification strategy. In sum, in each of these examples, causal assumptions provide the foundation for linking causal estimands to statistical estimands and thus for building an ADM model that allows for optimal decisions.

Why a causal lens is beneficial for reliability. We believe that adopting a causal perspective provides a systematic way to connect decision-making objectives, computational methods, and real-world deployment environments. This approach places causal identification at the center of ADM development. Hence, causal identification requires developers to clearly differentiate between causal estimands and statistical estimands, as well as to formalize when both are aligned. This shift in focus – from merely optimizing predictive accuracy to ensuring valid causal estimates – ensures that ADM systems can perform reliably in real-world settings.

A clear causal lens strengthens ADM systems by clarifying the assumptions required to ensure when and where decisions are reliable. Rather than seeing causal assumptions as a burden, we encourage to view them as a strength – a formal way to define the boundary conditions under which a system is designed to operate. Articulating these assumptions can also help in thinking about data requirements, such as identifying potential confounders that need to be measured. As such, transparency about causal assumptions not only strengthens trust when they hold but also clarifies when not to trust or deploy a model. Without causal reasoning, even seemingly well-performing models can fail in unpredictable ways, particularly in dynamic or complex environments. Nevertheless, we acknowledge that ADM models trained without a causal lens, or in contexts where assumptions are difficult to verify, may still perform well in certain cases. However, such models should be viewed as heuristics rather than robust solutions – which may be defendable for low-stakes tasks where causal reasoning is beneficial but not strictly needed.

So, what are the boundary conditions when a causal approach is mandatory, beneficial, or even harmful? A causal approach is generally mandatory in high-stakes scenarios where decisions directly affect safety, health, or legal outcomes. Examples include medical treatment recommendations or autonomous vehicle navigation, where understanding the causal effects of actions ensures reliability and mitigates risk. In such settings, unverifiable causal assumptions should serve as a 'red flag' for deploying ADM systems, as flawed causal reasoning can result in harmful consequences. Conversely, a causal lens is beneficial – but often not strictly necessary – in low-stakes or routine decisions, such as music recommendations or personalized marketing offers, where the cost of errors is minimal. Yet, a causal approach often leads to a better performance, and, hence, companies such as Spotify make broad use of causal approaches in their ADM tasks for business reasons[10]. However, in settings where causal relationships and the underlying assumptions are entirely speculative, causal claims may even be harmful by introducing unwarranted complexity, uncertainties, and even incorrect beliefs about a system's reliability. For instance, if an ADM system for marketing decisions mistakenly assumes that all relevant confounders (such as costumer motivation or seasonal trends) have been accounted for, it may falsely attribute increased sales entirely to a specific promotion, giving decision-makers undue confidence.

Another benefit of adopting a causal perspective in ADM is the ability to address transportability – that is, the transfer of causal results learned in one environment (for instance, a training setting) to another (for instance, a deployment setting). Concepts such as causal transportability[11] provide formal tools to specify the assumptions under which ADM systems can reliably generalize across settings. For example, a model trained to optimize traffic flow in one city may encounter differences in road infrastructure, weather patterns, or driver behavior when deployed in another. A causal approach can formalize which relationships remain invariant across these environments and help adjust the ADM system accordingly. Hence, transportability hinges on causality because causal relationships are robust to shifts between environments. Likewise, such a causal approach may help avoid shortcut learning – a common issue where models exploit spurious correlations or superficial patterns in the training data that may fail under deployment conditions. By making the causal assumptions explicit, developers can systematically assess whether the model is learning meaningful, transportable relationships or merely relying on shortcuts.

Challenges and directions for future research

Several practical challenges remain that must be addressed to ensure ADM systems operate effectively in dynamic, real-world environments.

Data quality. The reliability of any ADM system depends on the quality of the training data. Distribution shifts – meaning the data distribution during deployment differs from that during training – pose complex challenges, especially in dynamic settings where conditions change over time. While causal reasoning improves generalizability by identifying relationships that are invariant across environments, more research is needed to develop ADM models that can safely adapt to distribution shifts and are robust to external shocks and adversarial attacks. Ultimately this requires models that can accurately quantify their prediction confidence across different deployment conditions as well as novel monitoring techniques to assess ADM performance in dynamic data streams.

Data quality issues are multifaceted and can include measurement errors, representation biases, and incomplete data. These issues are well-studied in the area of governmental survey statistics, which has developed a rich toolkit for conceptualizing and assessing errors in data from a population inference perspective. Yet, data quality is not an absolute concept but rather depends on the specific application task. While survey science traditionally focuses on requirements for

valid descriptive inference, causal modeling hinges on coverage, that is, a non-zero inclusion probability of groups in the training data that differ in their treatment response. We thus call for research on new data audits, quality metrics, and validation frameworks that are tailored to the goals of (different classes of) ADMs, for which survey science provides valuable starting points.

Uncertainty quantification and robustness. Assessing the uncertainty in outcomes, decisions, and the overall model is critical for ensuring the reliability of ADM systems in practice. For example, in medical decision-making, uncertainty estimates can help determine whether the probability of a benefit from treatment is sufficiently large to outweigh the risks of adverse reactions. While substantial progress has been made in quantifying uncertainty for predictive machine learning, extending these methods to ADM settings presents new challenges such as additional uncertainty due to violations of identifiability assumptions or due to low treatment overlap. A causal perspective can help to formalize the different sources of uncertainty, that is, whether the uncertainty comes from a lack of identifiability of the causal estimand due to violations of causal assumptions (for example, unobserved confounding) or from estimatability issues such as low overlap (for example, certain individuals never receiving treatment) or lack of data.

A common challenge for ADM is to assess whether causal assumptions, such as the absence of unobserved confounding, hold in practice. Future research should focus on developing new methods that help in assessing the plausibility of assumptions and that account for potential violations (such as partial identification and causal sensitivity analysis). Another interesting route is to derive methods that optimize decisions over an uncertainty set of potentially confounded policies and that thus find worst-case guarantees in the form of confounding-robust policies[12]. In addition, there is a need for tailored methods that account for all those sources of uncertainty simultaneously. For example, there exist various methods for partial identification and sensitivity analysis that account for uncertainty due to lack of identifiability, but more research is needed on how to augment these methods to account for finite sample uncertainty (for instance, via conformal prediction or Bayesian inference). Together, this can yield more systematic frameworks for robust ADM systems under a range of causal assumptions, which will enable practitioners to better understand and mitigate risks associated with imperfect causal models.

To improve the robustness of ADM systems, another direction is the use of causal world models[13], which abstract underlying causal mechanisms to improve generalization and adaptation across domains and different decision-making objectives, such as optimizing under varying constraints or over different outcomes. Finally, many decision-making settings involve multiple objectives, so tailored methods for multi-criteria decision-making are needed that balance a primary objective while minimizing the risk of harm. For instance, in medicine, drugs must achieve high efficacy while ensuring safety with minimal side effects[5].

Performativity. Deploying an ADM system introduces performativity — the phenomenon where predictions made by a system actively shape the target of prediction[14]. Such a feedback loop between model and data once again points to the need for causal reasoning in ADM to explicitly model how predictions influence future outcomes. For instance, in smart cities, a traffic control system that optimizes signal timing to reduce congestion might influence drivers' route choices, leading to new traffic patterns that the system did not initially anticipate. Similarly, an AI agent may shape the behavior of humans, causing the data-generating process to be different from what the AI agent was originally trained on. Traditional predictive models fail to account for this, as they assume a stable data-generating process and do not capture how interventions — such as continued deployment of the system — reshape the future training data. A recent area of research, often framed under the concept of 'performative prediction', examines the causal influence of machine learning predictions on the very outcomes they aim to forecast. Various solution concepts and algorithms have been proposed, such as ensuring the stability of a system after repeated retraining. We encourage further research at the intersection of machine learning and causality — for example, using non-experimental data in settings where exploration is costly or restricted. Performative prediction also provides a valuable framework for studying more complex decision-making settings and strategic interactions, such as scenarios where multiple agents anticipate each other's actions and adapt their behavior accordingly. Such research will be especially important due to the growing adoption of AI agents in practice.

Evaluation and benchmarking. More research is needed to improve benchmarking in ADM systems, especially by developing new evaluation frameworks that reflect the complexities of real-world deployment settings. Since the estimands in ADM — such as counterfactual outcomes, treatment effects, or optimal actions — are generally unobservable, they cannot be traditionally benchmarked using standardized datasets. Rather, simulations and semi-synthetic datasets are typically the common approach to assess the performance of ADM systems[12]. Future work should prioritize the development of tailored benchmarks that incorporate realistic decision-making environments, accounting for dynamic conditions, feedback loops, and changing user behaviors. Promising directions for this include sequential A/B tests and bandit algorithms, which iteratively refine decision-making by adaptively allocating interventions based on observed outcomes, as well as adaptive clinical trials, which modify trial parameters in response to accumulating evidence to enhance efficiency and ethical considerations. These methods ensure benchmarks remain aligned with the evolving nature of decision environments. An alternative approach is to leverage real-world physical systems in a controlled environment as a testbed[15]. More research is also needed to construct better performance metrics that align closely with decision-making objectives, such as the quality of decision rules, downstream societal impacts, and performance across diverse subpopulations.

Conclusion

Causal reasoning is a necessary, but not sufficient condition for reliable decision-making in practice. ADM systems commonly involve deep interactions between algorithms and humans — raising critical questions such as non-compliance with algorithmic decisions and disparate impact downstream. Yet, causal reasoning fosters transparency about the conditions under which ADM systems can be trusted to operate reliably, providing an invaluable building block for safe and robust decision-making. Causal reasoning offers a powerful but also necessary foundation for improving the safety and reliability of ADM. By ensuring that ADM systems capture the true effects of decisions, causal assumptions help align the systems with real-world objectives.

Christoph Kern [1,2,3] ✉, **Unai Fischer-Abaigar** [1,2], **Jonas Schweisthal** [2,4], **Dennis Frauen** [2,4], **Rayid Ghani** [5],

Comment

Stefan Feuerriegel [1,2,4], **Mihaela van der Schaar** [6] & **Frauke Kreuter** [1,2,3]
[1]Department of Statistics, LMU Munich, Munich, Germany. [2]Munich Center for Machine Learning (MCML), Munich, Germany. [3]Joint Program in Survey Methodology (JPSM), University of Maryland, College Park, Maryland, USA. [4]LMU Munich School of Management, LMU Munich, Munich, Germany. [5]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Faculty of Mathematics, University of Cambridge, Cambridge, UK.
e-mail: christoph.kern@lmu.de

Published online: 27 May 2025

References

1. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806 (Association for Computing Machinery, 2017).
2. Dulac-Arnold, G. et al. *Mach. Learn.* **110**, 2419–2468 (2021).
3. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. *Science* **366**, 447–453 (2019).
4. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2009).
5. Feuerriegel, S. et al. *Nat. Med.* **30**, 958–968 (2024).
6. Manski, C. F. *Identification for Prediction and Decision* (Harvard Univ. Press, 2009).
7. Dawid, P. *J. Causal Inference* **9**, 39–77 (2021).
8. Coston, C., Mishler, A., Kennedy, E. H. & Chouldechova, A. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 582–593 (Association for Computing Machinery, 2020).
9. Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. *Am. Econ. Rev.* **105**, 491–495 (2015).
10. Heaven, W. D. The complex math of counterfactuals could help Spotify pick your next favorite song. *MIT Technology Review* (4 April 2023).
11. Bareinboim, E. & Pearl, *J. Proc. AAAI Conf. Artificial Intelligence* **26**, 698–704 (2021).
12. Kallus, N. & Zhou, A. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems 31 (NeurIPS)* (eds Bengio, S. et al.) (Curran Associates, Inc., 2018).
13. Richens, J. & Everitt, T. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations* (ICLR, 2024).
14. Perdomo, J., Zrnic, T., Mendler-Dünner, C. & Hardt, M. Performative prediction. In *Proceedings of the International Conference on Machine Learning* (eds Daumé, H. III & Singh, A.) 7599–7609 (PMLR, 2020).
15. Gamella, J. L., Peters, J. & Bühlmann, P. *Nat. Mach. Intell.* **7**, 107–118 (2025).

Acknowledgements

This work is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

Author contributions

Conceptualization and writing (original draft): C.K., U.F.-A., S.F.; writing (original draft): J.S., D.F.; supervision and writing (reviewing and editing): R.G., M.v.d.S., F.K.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Computational Science* thanks Stefan Lessmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

6 The Value of Prediction in Identifying the Worst-Off

Publication

U. Fischer-Abaigar, C. Kern, and J. C. Perdomo (2025). "The Value of Prediction in Identifying the Worst-Off". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. PMLR, pp. 17239–17261. URL: https://proceedings.mlr.press/v267/fischer-abaigar25a.html

Author Contributions UFA led the project as first author, developed the theoretical contributions building on prior work from JCP, ran all experiments and empirical analyses on the labor market data, created the visualizations, and wrote the original draft. JCP contributed to the initial conceptualization, supervised the project, and provided feedback and editing throughout. CK contributed to conceptualization, reviewing, and editing, and co-supervised the project.

The Value of Prediction in Identifying the Worst-Off

Unai Fischer-Abaigar [1 2] **Christoph Kern** [1 2] **Juan Carlos Perdomo** [3]

Abstract

Machine learning is increasingly used in government programs to identify and support the most vulnerable individuals, prioritizing assistance for those at greatest risk over optimizing aggregate outcomes. This paper examines the welfare impacts of prediction in equity-driven contexts, and how they compare to other policy levers, such as expanding bureaucratic capacity. Through mathematical models and a real-world case study on long-term unemployment amongst German residents, we develop a comprehensive understanding of the relative effectiveness of prediction in surfacing the worst-off. Our findings provide clear analytical frameworks and practical, data-driven tools that empower policymakers to make principled decisions when designing these systems.

1. Introduction

Faced with pressure to modernize, large bureaucracies are increasingly adopting risk prediction tools to improve efficiency and better serve their populations. Beyond optimizing aggregate outcomes, investments in these programs often aim to address historical inequities and prioritize the needs of the worst-off. For instance, in 2012, Wisconsin launched a risk prediction system to explicitly address deep racial disparities in academic achievement and improve high school graduation rates amongst underserved students. More broadly, such systems are particularly relevant in settings where normative considerations demand prioritizing those at the greatest risk of adverse outcomes, and where well-established downstream interventions can meaningfully benefit these vulnerable individuals.

From a design perspective, these risk predictors are challenging to evaluate because their value cannot be assessed without reference to the broader social context. The value of a risk predictor is ultimately determined by its impact on bottom-line welfare (e.g., graduation rates) and how these welfare impacts compare to those of other bureaucratic alternatives (Johnson & Zhang, 2022). For example, to understand whether investments in prediction are truly valuable in Wisconsin, we need to assess how much better the risk predictor is at identifying at-risk students relative to existing policies. We also need to understand whether sophisticated prediction systems yield higher graduation rates amongst the underserved than structural investments in teacher training or better facilities.

Equity-driven programs are pervasive in applications like social housing, poverty targeting, and unemployment assistance. In these contexts, many government agencies are exploring how algorithmic prediction systems may be an improvement over their current profiling processes (Körtner & Bonoli, 2023). Yet, due to the absence of an overarching framework that allows the systematic assessment of the relative impacts of different design decisions, efforts to improve predictive accuracy are rarely studied in concert with other policy levers such as expanding screening capacity.

Building on recent work in a budding area of learning in resource allocation contexts, we develop tools to evaluate the design and broader impact of prediction systems that aim to identify the worst-off members of a population. We develop a holistic understanding of the value of statistical prediction in these contexts through theoretical insights into foundational statistical models and a real-world case study on identifying long-term unemployment. Our results establish clear theoretical and empirical criteria characterizing the relative value of core design decisions within these problems. Specifically, we identify when improving prediction provides a higher marginal benefit in helping an institution identify the worst-off. This is compared to alternative strategies, such as keeping prediction accuracy fixed, expanding bureaucratic capacity and screening a larger population.

Interestingly, we show that prediction is a first and last-mile effort. The impacts of improving prediction are always outweighed by those of expanded screening capacity, except for when the system explains either none or almost all of the variance in outcomes. While this relationship is moderated by costs, it still largely holds when prediction improvements

[1]Department of Statistics, University of Munich (LMU), Munich, Germany [2]Munich Center for Machine Learning, Germany [3]Harvard University, Boston, MA, US. Correspondence to: Unai Fischer-Abaigar <Unai.FischerAbaigar@stat.uni-muenchen.de>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

are more cost-efficient than measures that expand access.

These results are counternarrative to current efforts in empirical public policy where agencies focus on incremental improvements within complex prediction systems, starting from the solid baseline performances of their current processes (Desiere et al., 2019; Desiere & Struyven, 2021). Furthermore, implementing more complex profiling systems at scale comes with operational costs (such as staff training and data collection) which need to be contextualized by the cost-benefit ratio of expanding access. Our empirical case study explicates how to systematically assess the relative gains of these design components in a real-world application setting, translating formal insights into critical guidance for designers of these systems.

Our results provide theoretically principled and empirically grounded tools for policymakers to make informed decisions when designing prediction systems to identify the worst-off. They also offer a practical framework to help determine how much should be invested in prediction relative to other interventions and how to decide when prediction systems are "good enough" for deployment.

1.1. Overview of Framework and Contributions

Setup. We consider a scenario where a decision-maker seeks to identify worst-off members of a population, as determined by a real-valued welfare metric $Y \in \mathbb{R}$, with the goal of prioritizing them for further screening and support. The population is represented by a distribution $\mathcal{D}$ over features X and outcomes Y. The planner aims to identify all individuals whose outcomes Y fall below some threshold $t(\beta)$, $Y \leqslant t(\beta)$. Here, $\beta \in [0, 1]$ is a parameter (quantile) that determines the size of the population that is at risk, $\Pr[Y \leqslant t(\beta)] = \beta$. For instance, in poverty prediction, Y is income, and the goal is to identify everyone whose income is below some value.

To solve this problem, the social planner has access to data $(X, Y) \sim \mathcal{D}$ and builds a screening policy $\pi : \mathcal{X} \to \{0, 1\}$ that determines whether an individual with features x is screened from the broader population to see if they belong to the worst-off group. Learning plays a fundamental role since the optimal policy is to predict each person's expected outcome, $f(x) = \hat{Y} \approx \mathbb{E}[Y \mid X = x]$ and screen those in the bottom fraction, $\pi_f(x) = 1\{f(x) \leqslant t(\alpha)\}$.

Unpacking this further, $\alpha \in [0, 1]$, is a design parameter that determines how many people the planner can screen, $\Pr[f(x) \leqslant t(\alpha)] = \alpha$. The amount of resources α need not be equal to the size of the target population β. For instance, an organization might have normative goal of identifying the poorest 5% of individuals, but only have the resource to screen 1% of the population. Conversely, they might realize that predictions are not perfect, and that to identify the bottom 5%, they might have to screen 10% of the population.

Given a predictor f, a screening budget of α, and a target parameter β, the value of a prediction system is equal to the fraction of the at-risk population that it identifies,

$$V(\alpha, f; \beta) = \Pr_{\mathcal{D}}[f(x) \leqslant t(\alpha) \mid Y \leqslant t(\beta)],$$

where again $t(\alpha), t(\beta)$ are chosen to respect the design constraints. We focus on this notion of value since our driving motivation is to analyze domains like unemployment assistance, or poverty prediction, where there is no harm in the prediction system raising a false positive ($\pi(x) = 1, Y > t(\beta)$). By and large, the true value of the system is equal to the extent that it helps an institution efficiently identify the needy amongst a large, diverse population.

The focus of our work is to build a holistic understanding of prediction in these contexts by evaluating the relative impacts of different design parameter, such as expanding screening capacity or improving prediction, on this notion of bottom-line value $V(\alpha, f; \beta)$. We develop these insights through theoretical investigations as well as in-depth empirical case study.

Mathematical Results. Following Perdomo (2024), we formalize the relative value of prediction for the worst-off by studying the *prediction-access ratio* or PAR. Intuitively, the PAR measures the relative change in value achieved by optimizing different policy levers.

$$\text{PAR} = \frac{\text{Marginal Value of Expanding Access}}{\text{Marginal Value of Better Prediction}}.$$

We formally define this quantity in Equation 3. While initially developed to specifically study the value of prediction in allocation problems where allocating goods to individuals had heterogeneous effects, here we extend this concept to analyze the value of prediction in a related, but distinct, setting where we aim to identify the worst-off.

Small values of the PAR (i.e. PAR < 1) indicate that small improvements in prediction yield a much larger (relative) impact in the ability to target the worst-off than a small expansion in screening capacity. The opposite is true if the PAR is greater than 1. Calculating this quantity is a fundamental step in deciding which policy lever makes economic sense.

Costs. A full cost-benefit analysis requires combining the prediction-access ratio with the (marginal) costs of improvements in capacity C_{Access} and prediction C_{Pred}. Once we factor in costs, it is easy to decide what to focus on. A social planner should expand access whenever

$$\frac{C_{\text{Access}}}{C_{\text{Pred}}} < \text{PAR}$$

and invest in better prediction otherwise. The (marginal) costs that competing policy levers carry are inherently

context-dependent and will vary across application domains. In many applied settings the cost ratio is comparatively well understood, for example, the salary of an additional caseworker or the cost of a household survey. By presenting PAR separately from costs, we isolate the welfare side of the equation; domain experts can then plug in their own cost estimates to reach a policy decision. In particular, the PAR tells us how much we should be willing to *pay* for improvements in prediction versus expanding access.

We encourage future work to explore scenarios with more complex or less clearly defined cost structures. For instance, many practical applications involve recurring costs, such as ongoing staff salaries or periodic data collection, and fixed costs, such as initial investments in infrastructure or predictive model development. Analyzing how these cost structures affect welfare decisions over time, including amortization of fixed investments or identifying the point at which specific improvements become cost-effective, would significantly enhance our understanding of the relative value of prediction.

To build intuition for the value of prediction in identifying the worst-off, we examine the prediction access ratio in one of the most basic statistical models. The outcomes Y are Gaussian, and the learner has access to a predictor $f(x) = \hat{Y}$ such that the errors $Y - \hat{Y}$ are also Gaussian and independent of $\hat{Y}$. While extremely simple, the model yields surprisingly precise numerical insights that exactly match up in our real-world case study, where, of course, none of these assumptions hold. In this setting the quality of $\hat{Y}$ is fully summarized by the coefficient of determination $R^2 = \mathrm{corr}(Y, \hat{Y})^2$.

Our first result identifies when local improvements in prediction have the highest impact:

Theorem 1.1 (Informal, see Theorem 3.2)**.** *If α is at least a constant, the local improvements in V with respect to R^2 diverge in two regimes: (1) $R^2 \to 1$ and $\alpha = \beta$, or (2) $R^2 \to 0$. In both cases, the prediction-access ratio satisfies* $\mathrm{PAR}(\alpha, \beta) = 0$.

Predictions have the highest marginal impact at low and high R^2-values, making them a first- and last-mile effort. Our second result characterizes when the opposite is true. We prove that whenever screening capacities are severely limited relative to the size of the population one aims to identify $\alpha \ll \beta$, the benefits of increasing α are overwhelming. Furthermore, it shows that the impacts of improving access are still relatively larger exactly in the regime where most current systems operate: f explains $\approx 20\%$ of the variance and α is equal to, or even slightly larger, than β.

Theorem 1.2 (Informal, see Theorem 3.1, Proposition 3.3)**.** *If the predictor f explains an R^2 fraction of the variance, where R^2 is at least a constant, then the prediction access ratio is at least $\Omega(\alpha^{-1/(1-R^2)})$. Furthermore, if $0.15 \leqslant R^2 \leqslant 0.85$ and $\alpha \leqslant \beta$ or $0.2 \leqslant R^2 \leqslant 0.5$, $\beta \geqslant 0.15$, and $\alpha \leqslant 0.5$ then the local prediction-access ratio is at least 1.*

Empirical Results. We complement our theoretical discussion by presenting a methodology for policymakers to evaluate the prediction-access ratio in practice. Using a real-world administrative dataset on hundreds of thousands of jobseekers in Germany, we show that our theoretical findings generalize to a more complex, real-world context that closely resembles algorithmic profiling systems widely implemented in many countries. Notably, our results reveal that when considering non-local improvements, expanding screening capacity has an even greater impact compared to enhancing prediction accuracy.

1.2. Related Work

Machine learning is increasingly used in the public sector to allocate support by predicting individuals at risk of adverse outcomes (Fischer-Abaigar et al., 2024), with applications spanning a wide range of problem domains (Desiere et al., 2019; Blumenstock, 2016; Perdomo et al., 2023; Chan et al., 2012; Potash et al., 2015; Chouldechova et al., 2018). A large methodological literature draws on decision theory, operations research, economics, and machine learning to learn allocation rules from data (Elmachtoub & Grigas, 2022; Kitagawa & Tetenov, 2018; Manski, 2004; Fernández-Loría & Provost, 2022), with recent work in causal inference focusing on learning treatment policies from observational data (Athey & Wager, 2021; Kallus, 2021). However, many decision-makers rely on separately trained predictive risk scoring-systems to solve "prediction policy problems" (Kleinberg et al., 2015). Recently, this work has been extended using causal inference to train and evaluate these systems (Coston et al., 2023; Guerdan et al., 2023; Boehmer et al., 2024).

The widespread use of risk-scoring systems has raised concerns regarding their tradeoffs, pitfalls, and validity (Wang et al., 2024; Coston et al., 2023; Fischer-Abaigar et al., 2024). These concerns include not only questions of empirical performance but also of fairness and equity in how predictive systems shape access to public services (Barocas et al., 2023). Recent work explores alternative design choices—such as employing aggregate rather than individual-level predictions (Shirali et al., 2024), balancing immediate needs with information-gathering (Wilder & Welle, 2024), and introducing randomization (Jain et al., 2024)—to improve downstream outcomes.

Perdomo (2024) studies the prediction-access ratio under both linear and probit models, with the latter closely related to our work. While they focus on binary welfare outcomes, we adopt a continuous welfare metric and a distinct policy objective: rather than evaluating changes in overall expected

welfare, we measure the fraction of truly worst-off individuals who are identified. This captures a mathematically and conceptually distinct setting frequently encountered in the public sector. For instance, employment agencies often prioritize identifying and assisting individuals in greatest need, rather than optimizing average employment outcomes across all jobseekers. In addition, we introduce a set of empirical tools to analyze these tradeoffs in practice, while the work of Perdomo (2024) is purely theoretical.

2. Formal Framework

We start by formally defining our screening problem.

Definition 2.1 (Screening Problem)**.** The screening problem seeks to identify a decision rule $\pi\colon \mathbb{R} \to \{0,1\}$ that fraction of the worst-off population that is identified while adhering to resource constraints $\alpha \in (0,1)$ that bound the percentage of the population that can be screened by the social planner:

$$\max_{\pi\colon \mathbb{R}\to\{0,1\}} \mathbb{E}\left[\pi(\hat{Y}) = 1 \mid Y \leqslant F_Y^{-1}(\beta)\right] \text{ s.t. } \mathbb{E}\left[\pi(\hat{Y})\right] \leqslant \alpha$$

The quantile $F_Y^{-1}(\beta)$ denotes the welfare cutoff that identifies the worst-off $\beta \in (0,1)$ fraction of the population.

Given perfect knowledge of the welfare outcomes $\hat{Y} = Y$, the optimal decision policy is simple: rank individuals based on their outcomes Y and intervene in the bottom α-fraction of the population. In the general case, we have:

Proposition 2.2. *The optimal policy $\pi^*\colon \mathbb{R} \to \{0,1\}$ to solve the screening problem (Definition 2.1) is equal to $\pi^*(\hat{Y}_i) = 1\{s(\hat{Y}_i) \geqslant F_s^{-1}(1-\alpha)\}$ where $F_s^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of $s(\hat{Y}) = \Pr[Y \leqslant F_Y^{-1}(\beta) \mid \hat{Y}]$.*

Policy Value in Gaussian Setting. For the theoretical investigation, we assume independent, identically distributed errors $\varepsilon = Y - \hat{Y} \overset{iid}{\sim} \mathcal{N}(0,\gamma^2)$ that are independent of $\hat{Y}$. In this setting, the screening problem can be solved by ranking individuals in ascending order of their predicted outcomes $\hat{Y}$ and screening the bottom α-fraction (see Proposition C.1), achieving the policy value:

$$V(\pi^*) = \Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y \leqslant F_Y^{-1}(\beta)] \tag{1}$$

In addition, we assume welfare outcomes $Y \sim \mathcal{N}(\mu,\eta^2)$. Because ε is independent of $\hat{Y}$, this implies that Y and $\hat{Y}$ follow a bivariate normal distribution.

Proposition 2.3. *(Policy Value in Gaussian Setting) Let $Y - \hat{Y} \overset{iid}{\sim} \mathcal{N}(0,\gamma^2)$ and $Y \sim \mathcal{N}(\mu,\eta^2)$, then the value $V(\pi^*)$ of the optimal screening policy π^* is given by*

$$V(\pi^*) = V(\alpha,\beta,R^2) = \frac{\Phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)}{\beta} \tag{2}$$

where $\Phi_2(\cdot)$ denotes the bivariate standard normal CDF with correlation $\rho = \sqrt{\eta^2-\gamma^2}/\eta$ and $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution.

In this model, the goodness of the predictions $\hat{Y}$ are entirely captured by the coefficient of determination R^2, which equals the squared correlation ρ^2 between Y and $\hat{Y}$.

Our analysis extends to the log-normal distribution $\log Y \sim \mathcal{N}(\mu,\eta^2)$ under a a multiplicative error model $Y = \hat{Y}\cdot u$ with $\log u \sim \mathcal{N}(0,\gamma^2)$. Taking logarithms, leads to $\log Y = \log \hat{Y} + \log u$. Since the logarithm is strictly increasing, the ordering of Y and $\hat{Y}$ is preserved under transformation. This allows us to apply the same framework to the log-transformed variables $\log Y$ and $\log \hat{Y}$. This extension is particularly useful because many welfare outcomes, such as income distributions (Clementi & Gallegati, 2005), can be approximated by a log-normal distribution.

Visualization. For a given screening capacity α and R^2 value, we can illustrate the corresponding screening policy by plotting the probability $\Pr\{\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y = y\}$ that an individual with welfare outcome $Y = y$ is screened. As shown in Figure 1, lower values of Y correspond to higher probabilities of being screened. We focus on evaluating how effectively the screening policy identifies individuals in the worst-off segment of the population (i.e., on the left side of the β cutoff).

3. Theoretical Results

The decision-maker has (at least) two pathways to raise the policy value, which we refer to as *policy levers*:

- **Expanding Access** Increasing the screening threshold from α to $\alpha + \Delta_\alpha$. If full screening were possible ($\alpha = 1$), the β-fraction would be fully identified, as $V(\pi^*) = \frac{\Phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)}{\beta} = \frac{\Phi\left(\Phi^{-1}(\beta)\right)}{\beta} = 1$.
- **Improving Predictions** Investing in better predictive models, modeled as increasing R^2 to $R^2 + \Delta_{R^2}$. Perfect predictions ($R^2 = 1$) leads to optimal allocation of available capacities: $V(\pi^*) = \frac{1}{\beta}\Phi\left(\min(\Phi^{-1}(\alpha), \Phi^{-1}(\beta))\right)$.

Figure 1 showcases improvements in access and prediction. Increasing capacity expands the fraction of the population screened, while improving R^2 shifts probability mass across the β threshold, enhancing targeting accuracy.

Following Perdomo (2024), a key quantity of interest is the *prediction-access ratio* (PAR), which quantifies the relative improvements in policy value from enhancing predictions versus improving access to screening. Specifically, the PAR is defined as:

$$\text{PAR} = \frac{V(\alpha+\Delta_\alpha,\beta,R^2) - V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta_{R^2}) - V(\alpha,\beta,R^2)} \tag{3}$$

In other words, the PAR can inform a social planner how much more they should be willing to pay for improvements in screening capacity relative to prediction. For example, a PAR > 2 implies that expanding the screening capacity by

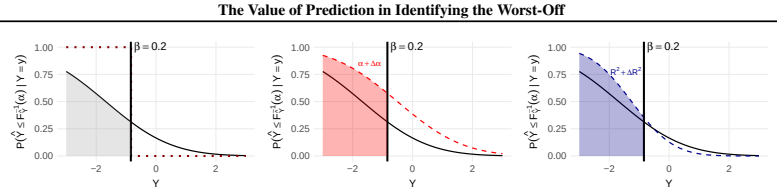


Figure 1. **Screening Policy in Gaussian Setting.** (Left) Probability of being screened for an individual with a specific welfare outcome Y, given $R^2 = 0.25$, $\alpha = 0.2$, and $\beta = 0.2$. The dashed line represents the unconstrained oracle policy, which perfectly screens those in need. (Middle) Policy with expanded screening capacity, where α increases by $\Delta_\alpha = 0.2$. (Right) Policy under an improved prediction model with $R^2 + \Delta_{R^2}$, where $\Delta_{R^2} = 0.2$. The shaded area under $\Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y = y]$, weighted by $f_Y(y)$ and normalized by $\Pr[Y \leqslant F_Y^{-1}(\beta)]$, corresponds to the policy value.

Δ_α yields at least twice the increase in policy value compared to investing in improved predictions by Δ_{R^2}. Consequently, the social planner should prioritize investments in screening capacity, provided the costs of doing so are not more than double those of improving predictions.

3.1. Theoretical Bounds for the Prediction-Access Ratio

In our setting, direct calculation of the PAR is challenging due to the policy value being analytically intractable and the problem featuring strong non-linearities. We derive bounds for specific cases and regimes that we consider particularly insightful, with a focus on marginal local improvements. In our empirical investigation, we find that the main results generalize well to a more complex, real-world setting.

What should priorities be if screening is very limited?

Theorem 3.1 (PAR for Small Screening Capacities)**.** *For any $0 < R^2 < 1$, $\Delta_{R^2}, \Delta_\alpha > 0$ and $0 < \beta \leqslant 0.5$ there exists a threshold $t(\beta, R^2, \Delta_{R^2})$ such that for any $\alpha + \Delta_\alpha \leqslant t$,* $\text{PAR}(\alpha, R^2, \Delta_\alpha, \Delta_{R^2})$ *is at least*

$$\frac{\Delta_\alpha}{\Delta_{R^2}} \sqrt{R^2(1-R^2)} \left(5.1 \cdot \alpha \Phi^{-1}(1-\alpha)\right)^{-\frac{1}{1-R^2}+o(1)}$$

where $o(1)$ goes to zero as α approaches zero.

Suppose the available screening capacity $\alpha + \Delta_\alpha$ is very small ($\alpha + \Delta_\alpha \ll \beta$), and assume there is a baseline level of predictability (i.e., R^2 is bounded away from 0). Then Theorem 3.1 implies that the PAR can become very large. Specifically, for small α, $\Phi^{-1}(1-\alpha)$ grows asymptotically like $\sqrt{\log(1/\alpha)}$. Consequently, the polynomial growth of $\alpha^{-1/(1-R^2)}$ drives the PAR to increase rapidly as α decreases. It follows that in the scarce capacity regime, expanding the screening capacity has a far greater impact than improvements in prediction accuracy.

When does prediction have the highest impact?

Theorem 3.2 (Maximally Effective (Local) Prediction Improvements)**.** *Let $0 < \beta < 1$ be fixed and $0 < \alpha < 1$. Consider the points that maximize the local rate of change in policy value V with respect to improvements in R^2:*

$$(\alpha_*, R^2_*) = \underset{(\alpha, R^2) \in (0,1)\times(0,1)}{\arg\max} \lim_{\Delta \to 0} \frac{V(\alpha, \beta, R^2+\Delta) - V(\alpha, \beta, R^2)}{\Delta}$$

The local improvements in V diverge—and are maximized—in two regimes: (1) $R^2_* \to 1$, $\alpha_* = \beta$, and (2) $R^2_* \to 0$. For both regimes, setting $\Delta_{R^2} = \Delta_\alpha = \Delta$, the local prediction-access ratio satisfies $\lim_{\Delta \to 0} \text{PAR}(\alpha, \beta, \Delta) \to 0$.

According to Theorem 3.2, marginal improvements in prediction are most impactful in two distinct regimes. First, when predictive capacity is very low, even a small initial investment can lead to disproportionately large improvements, provided that a minimal baseline of screening capacity is present. Second, as R^2 approaches one, further marginal improvements can also have a significant relative impact, specifically around the point where the screening capacity α matches the requirements for screening the entire β-segment of the population. See Figure 2.

When are small increases in screening capacity more impactful than improving predictions?

Proposition 3.3 (PAR for Local Improvements)**.** *Let R^2, β, and α satisfy either $R^2 \in (0.15, 0.85)$, $\beta \in (0.03, 0.5)$, and $\alpha \leqslant \beta$, or $R^2 \in (0.2, 0.5)$, $\beta \geqslant 0.15$, and $\alpha \leqslant 0.5$. If $\Delta_{R^2} = \Delta_\alpha = \Delta$, then* $\lim_{\Delta \to 0} \text{PAR}(\alpha, \beta, \Delta) \geqslant 1$.

We find that the PAR remains above one as long as $\alpha \leqslant \beta$ and R^2 is not too extreme. For larger β values (i.e., $\beta \geqslant 0.15$) the PAR stays above one even for large α provided R^2 remains in a moderate range. Crucially, this represents the standard parameter regime in which most allocation programs operate, characterized by a moderate baseline of predictions and resource levels comparable to β.

Numerical Simulations. We complement our theoretical investigation with numerical simulations of the PAR for different α, β and R^2 values (see Figure 2). Consistent with our theoretical results, the PAR becomes large for small screening capacities ($\alpha \ll \beta$) and remains above one for $\alpha \leqslant \beta$, provided a small baseline level of predictive performance has been established. The bounds in Proposition 3.3

are conservative, with PAR > 1 observed for a broad range of R^2 values. Prediction improvements are particularly impactful when R^2 is small. Although the PAR falls below one in the high-R^2 and high-α regime, allocation is nearly perfect, making further improvements a "last mile" effort.

Discussion. We found several insights relevant to policymakers aiming to iteratively improve a screening system. First, establishing a baseline level of predictive performance is usually a good starting point. Once this is achieved, expanding the screening capacity becomes the next priority. For very small capacities, Theorem 3.1 tell us that the PAR can increase significantly, making investments in screening capacity highly impactful.

Generally, expanding capacity to at least the level where everyone in need could hypothetically be screened ($\alpha \geqslant \beta$) is likely cost-efficient. Once both screening capacity and predictive accuracy are high and the allocation system is close to optimal, improvements in prediction become relatively more valuable again for perfecting the system. However, this regime may rarely be reached in practice.

In Figure 2, we display the PAR for a cost ratio of $1/4$. As expected, the regions where investing in R^2 is more efficient expand, and some of the earlier nonquantitative bounds no longer apply. Nevertheless, the key insights remain consistent: when screening capacities are small, investments in expanding them are very effective, while improvements in R^2 are more important when predictive accuracy is low.

4. Empirically Evaluating the PAR

While our theory offers broad intuition when expanding screening capacity or improving predictions is most effective, policymakers need practical tools for their own systems. To support this, we develop a methodology to compute and interpret the prediction-access ratio using empirical data, helping social planners identify the most efficient policy levers for their unique problem context.

Policy Value. As before, we define the allocation policy's value as the probability that the worst-off individuals are successfully identified:, i.e. $V(\alpha, \beta) = \Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y \leqslant F_Y^{-1}(\beta)]$. In practice, this can be measured using a recall-like metric, capturing the proportion of truly at-risk individuals screened by the policy.

$$V(\alpha, \beta) \approx \frac{\sum_{i=1}^{n} 1\{\hat{Y}_i \leqslant F_{\hat{Y},n}^{-1}(\alpha)\} 1\{Y_i \leqslant F_{Y,n}^{-1}(\beta)\}}{\sum_{i=1}^{n} 1\{Y_i \leqslant F_{Y,n}^{-1}(\beta)\}}$$

Increasing Screening Capacity. Given a chosen Δ_α the policy improvement can be directly computed $V(\alpha + \Delta_\alpha, \beta) - V(\alpha, \beta)$ by recalculating the empirical policy value at the new threshold. For example, in cash transfer programs (Blumenstock, 2016), a key question is how many resources α^* are required to reach a specified fraction p of poor households, i.e. $\alpha^* = \inf_{\alpha \in (0,1)} \{\alpha \colon V(\alpha, \beta) \geqslant p\}$.

Improving Predictions. A decision-maker can improve a model's predictions through various pathways:

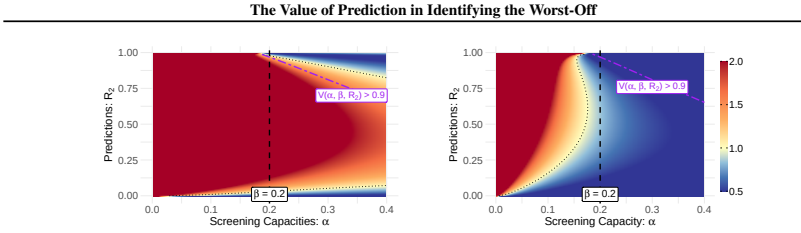
a) **Data Collection** Collect additional samples and increase the frequency of data collection. Social prediction systems are often vulnerable to distribution shifts over time in dynamic and evolving environments (Fischer-Abaigar et al., 2024; Aiken et al., 2023).
b) **Data Quality** Improve data quality (i.e., reduce errors and missing data) by means such as standardizing data collection processes, implementing centralized data management systems, and offering targeted training programs for staff.
c) **Collect Additional Features** In government, this may involve integrating separate data sources across institutions (Sun & Medaglia, 2019; Wirtz et al., 2019).
d) **Advanced Modeling Techniques** Utilize more sophisticated modeling techniques, which might capture more complex patterns in the data but are often more costly to operationalize.

In resource-constrained settings, planners often focus on incremental improvements rather than rebuilding entire systems. For instance, collecting a small amount of additional data may boost R^2 by a few points, uniformly reducing errors. To simulate such minor gains, we scale the model's residuals $\hat{Y}_+ = \hat{Y} + \delta(Y - \hat{Y})$, choosing $\delta \in (0, 1)$ so that R^2 increases by a target Δ_{R^2} (see Appendix B.3). This preserves the overall error structure, allowing us to gauge how a "similar but slightly better" model affects policy outcomes.

This approach can be extended in several ways. For example, residuals could be adjusted for specific subgroups to account for uneven prediction improvements (e.g., targeted data collection for rural or underrepresented populations). Alternatively, planners could retrain models under different conditions—such as sample size, feature set, or architecture—and compare the resulting policy value.

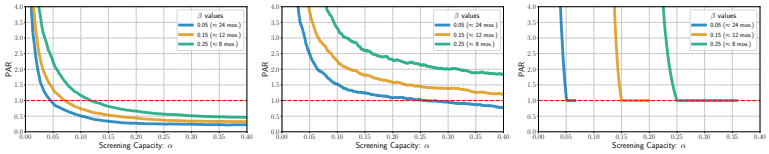
5. Case Study: Identifying Long-Term Unemployment in Germany

Public employment services (PES) across the globe make use of profiling approaches to identify jobseekers at risk of long-term unemployment to target preventative measures (Loxha & Morgandi, 2014). Starting from traditional rule-based approaches, many PES either test or already deploy algorithmic profiling to identify jobseekers in need of support (Desiere et al., 2019; Körtner & Bonoli, 2023). While these profiling tools assist in allocating programs that account for large shares of PES spending — making design choices critical (Kern et al., 2024) — systematic assessments of their relative value compared to other measures for



(a) PAR (b) $1/4 \times$ PAR

Figure 2. **Numerical Simulation of the Prediction-Access Ratio (PAR)**, Equation 3, for $\Delta_{R^2} = \Delta_\alpha = 0.01$ and $\beta = 0.2$. (Left) The PAR values. (Right) $1/4 \times$ PAR, representing a cost ratio of $1/4$. Each point represents a screening capacity α (x-axis) and R^2 value (y-axis), with the color bar showing the PAR clipped to the range $[0.5, 2.0]$. Dotted black lines represent PAR $= 1$, where improvements in α and R^2 are equally effective. The purple line marks the region in the (α, R^2) space where the policy value $V(\alpha, \beta, R^2)$ exceeds 0.9.



(a) $R^2 = 0$ (b) $R^2 = 0.15$ (c) $R^2 = 0.9$

Figure 3. **Prediction-Access Ratio** for $\Delta_{R^2} = \Delta_\alpha = 0.1$ across three regimes: (left) constant prediction ($R^2 = 0$), (middle) trained model ($R^2 = 0.15$), and (right) near-perfect prediction ($R^2 = 0.9$). As expected from our theoretical intuition, the PAR is large for small α and in the middle plot, which represents the typical regime for allocation systems.

improving jobseekers' outcomes remain absent.

We secured access to a dataset[1] on German jobseekers derived from German administrative labor market records that cover a large portion of the German labor force. It merges multiple administrative data sources, containing a wide spectrum of individual labor market information — including records on employment histories, received benefits, unemployment periods, participation in job training programs and demographic information. Such administrative records are the primary data source used by PES to build algorithmic profiling models (Bach et al., 2023).

Experimental Setup. We train a model to predict how long a newly registered jobseeker remains unemployed, defining the target Y as unemployment duration in days (capped at 24 months). Following Bach et al. (2023), we use a set of covariates capturing demographic information, labor market history, and most recent job details. To ensure full 24-month observations and mimic a realistic deployment scenario, we focus on unemployment spells beginning between 2010 and 2015, resulting in data on 274,515 different jobseekers and 553,980 unemployment spells. We refer to Appendix B.1 for additional information on the experimental setup and data.

Our focus is the β-fraction of jobseekers with the longest expected unemployment durations, representing those most at risk. In Germany, being unemployed for over one year (about 15% of cases in our data; Figure 8) meets the legal definition of long-term unemployment (Bach et al., 2023), but some countries adopt different cutoffs (Desiere et al., 2019).

5.1. Results

We train a `CatBoost` model (see Appendix B.2 for details), achieving an R^2 of 0.15 on the test set. This level of predictive power aligns well with what is typically observed in social prediction tasks (Salganik et al., 2020) and similar applied settings (Desiere et al., 2019).

How much does the screening capacity need to increase to target a significant fraction of high-risk jobseekers? As expected, larger screening capacities increase both the policy value and the number of high-risk jobseekers screened (see Figure 4(a)). Focusing on the (German) LTU

[1]The dataset is provided via a Scientific Use File by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) (Schmucker & vom Berge, 2023a;b).

cutoff ($\beta \approx 0.15$), our policy value aligns well with findings of previous studies[2] (Bach et al., 2023).

A planner might begin by setting $\alpha = \beta$, ensuring that, in theory, enough capacity is provided to screen and support every high-risk jobseeker. A natural question then arises: how much additional capacity Δ_α would be required to screen at least a specified percentage of high-risk individuals? This additional capacity represents the overhead that must be invested to account for imperfect predictions. We observe that the Δ_α required to ensure at least 75% of high-risk jobseekers are screened remains consistently around 0.25 across different β values. While the policy value increases as $\alpha = \beta$ rises, the marginal improvements gained from increasing access decrease for higher α, resulting in a somewhat stable Δ_α across β. In practice, this means we need to screen 25% more of the population to ensure adequate coverage.

What is the impact of improving screening capacity versus prediction errors? We simulate small improvements in the R^2 value by uniformly scaling the residuals by a multiplicative factor. To ensure that this approach approximates a realistic pathway of (marginally) improving the model, we train various models at different sample sizes. We then verify that as R^2 increases with the amount of training data, the variance of the residuals decreases, while the distribution remains largely unchanged in shape (see Figure 12). We then evaluate the prediction-access ratio for $\Delta_{R^2} = \Delta_\alpha = 0.1$ in three scenarios : (1) the trained `CatBoost` model with $R^2 = 0.15$, (2) near-perfect predictions with $R^2 = 1 - \Delta_{R^2}$ and (3) constant predictions ($R^2 = 0$), effectively randomizing screening decisions.

We observe a rise in the PAR for small screening capacities α (see Figure 3), consistent with Theorem 3.1. Under random allocation ($R^2 = 0$), the PAR stays below one for $\alpha \geqslant 0.1$. This result aligns somewhat with Theorem 3.2, where we found that the (local) PAR approaches zero as $R^2 \to 0$. Because we consider $\Delta = 0.1$ (rather than an infinitesimal improvement, see Figure 13 for $\Delta = 0.01$), the PAR remains large at small α. For the `CatBoost` model ($R^2 = 0.15$), capacity improvements stay relatively more effective (i.e., PAR > 1) for larger α, matching Proposition 3.3, where we found that for moderate R^2 and $\alpha \leqslant \beta$, the local PAR remains above one. Meanwhile, near-perfect predictions ($R^2 = 0.9$) make capacity investments highly efficient, causing the PAR to diverge for $\alpha < \beta$, then drop sharply near $\alpha = \beta$ because the allocation becomes nearly optimal. When $\alpha \geqslant \beta$, the PAR stabilizes at one as numerator and denominator both approach zero.

These observations broadly match our theoretical findings, despite the non-local improvements and more complex residual structure. Notably, the theory's focus on local improvements offers a conservative perspective on capacity investments: even under random allocation ($R^2 = 0$), securing a modest screening capacity ($5-10\%$) is often the first priority, while at very high R^2, gains in policy value diminish so rapidly once $\alpha \geqslant \beta$ that the relative advantage of further prediction investments becomes negligible.

When do small improvements in prediction error have the largest impact? From theory (Theorem 3.2), we expect local policy value improvements from better predictions to diverge as $R^2 \to 0$ and $R^2 \to 1$ when $\alpha = \beta$. This aligns with our results in Figure 4: for small Δ_{R^2}, the rate of local improvements in $V(R^2)$ with respect to R^2 diverges. The location of the maximum in α also follows from the theory: as $R^2 \to 1$, the rate only diverges for $\alpha = \beta$, while for small R^2 the maximum is at $\alpha \approx 0.5$.

What are the relative benefits and tradeoffs of using a simpler vs more complex prediction model? We compare a shallow 4-depth decision tree with the `CatBoost` model. As expected, the simpler tree shows a small drop in predictive power (5% decrease in R^2) which translates into a 1–8% reduction in policy value (see Figure 15). Compared to a uniform 5% increase in R^2 achieved by scaling the residuals (see Figure 14), the differences in policy value are only partially similar across α. The `CatBoost` model does not provide a uniform improvement over the decision tree; for instance, it performs better at distinguishing longer unemployment spells.

Despite this performance gap, the simpler model offers potential advantages: it fits on a single sheet of paper, demands minimal computational infrastructure, can be easily explained to frontline case workers and resembles the categorical prioritization rules common in public institutions. (Johnson & Zhang, 2022). Because more complex models incur higher costs, a planner might instead increase screening capacity. Formally, we define

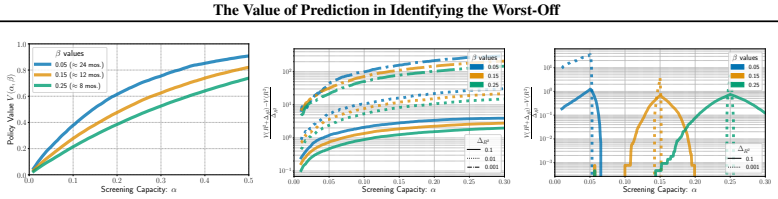
$$\Delta^*_\alpha = \inf_{\Delta_\alpha \in (0, 1-\beta)} \left\{ \Delta_\alpha : \frac{V_{\text{TREE}}(\alpha+\Delta_\alpha, \beta) - V_{\text{TREE}}(\alpha, \beta)}{V_{\text{CAT}}(\alpha, \beta) - V_{\text{TREE}}(\alpha, \beta)} \geqslant 1 \right\}$$

the smallest Δ^*_α that matches the policy-value gains of the `CatBoost` model. Empirically, Δ^*_α mostly rises with α (see Figure 15), consistent with our finding that the PAR decreases with α. By framing the difference between models in terms of additional screenings, planners can directly compare the cost of increased capacity to that of deploying a more complex model.

6. Conclusion

This paper develops a framework for quantifying the relative value of prediction in identifying the worst-off. We formalize tradeoffs between expanding screening capacity and

[2] For the percentage of correctly identified LTU episodes, they report values of 0.29 at $\alpha \approx 0.1$ and 0.58 at $\alpha \approx 0.25$, compared to our observed values of 0.28 and 0.56, respectively.



(a) Policy Value (b) $R^2 = 0$ (c) $R^2 = 1 - \Delta_{R^2}$

Figure 4. (a) Policy value across different screening capacities (α) and worst-off fractions β evaluated on the test set using the `CatBoost` regression model. A β value of 0.15 corresponds to the 12-month cutoff used to define long-term unemployment in Germany. (b, c) The rate of local improvements in $V(R^2)$ with respect to small changes in R^2. Panel (b) shows that the local improvements become increasingly large as Δ_{R^2} approaches zero. Panel (c) illustrates that improvements in prediction have the greatest impact when the capacity precisely matches the targeted fraction of the population ($\alpha = \beta$). Note that these are on a logarithmic scale.

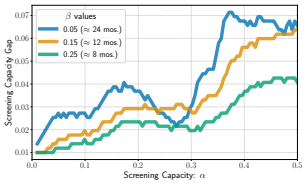


Figure 5. The minimum additional screening capacity that would need to be invested for the decision tree to achieve a policy value comparable to that of the `CatBoost` model.

improving predictive models, and show through both mathematical analysis and a real-world case study that prediction is not always the most important piece of the puzzle in social allocation systems. Future work could examine more specific application settings and cost structures, including distinctions between fixed and recurring costs, and explore policy levers that improve prediction unevenly, for example, by reducing errors in high-risk subgroups or by increasing robustness to distributional shifts. More broadly, we see a need for clearer theoretical foundations to understand the role of prediction in public-sector allocation, particularly in relation to the institutional and administrative systems in which it is embedded.

Acknowledgements

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research and by the Volkswagen Foundation, grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)". Juan Carlos Perdomo is supported by the Center for Research on Computation and Society (CRCS) at Harvard University and by the Alfred P. Sloan Foundation grant G-2020-13941. We would like to thank the anonymous reviewers for their insightful comments, as well as Frauke Kreuter, Patrick Schenk and Moritz Hardt for their valuable feedback.

Impact Statement

Our work offers a principled framework for evaluating the relative benefits of using predictive models to target the most vulnerable populations, helping public agencies allocate limited resources more effectively. However, formalizing complex institutional processes inevitably omits important real-world details, risking biases or misalignments if assumptions are not carefully examined. We encourage policymakers and researchers to incorporate fairness, transparency, and accountability measures when implementing these methods, particularly in resource-constrained contexts where small design changes can disproportionately affect marginalized communities.

References

Aiken, E., Ohlenburg, T., and Blumenstock, J. Moving targets: When does a poverty prediction model need to be updated? In *Proceedings of the 6th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '23, pp. 117, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701498. doi: 10.1145/3588001.3609369. URL https://doi.org/10.1145/3588001.3609369.

Athey, S. and Wager, S. Policy Learning with Observational Data. *Econometrica*, 89(1):133–161, 2021.

Bach, R. L., Kern, C., Mautner, H., and Kreuter, F. The impact of modeling decisions in statistical profiling. *Data & Policy*, 5:e32, 2023. doi: 10.1017/dap.2023.29.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Blumenstock, J. E. Fighting Poverty with Data. *Science*, 2016.

Boehmer, N., Nair, Y., Shah, S., Janson, L., Taneja, A., and Tambe, M. Evaluating the Effectiveness of Index-Based Treatment Allocation. *arXiv preprint arXiv:2402.11771*, 2024.

Chan, C. W., Farias, V. F., Bambos, N., and Escobar, G. J. Optimizing Intensive Care Unit Discharge Decisions with Patient Readmissions. *Operations Research*, 60(6):1323–1341, 2012. doi: 10.1287/opre.1120.1105. URL https://doi.org/10.1287/opre.1120.1105.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148. PMLR, 2018.

Clementi, F. and Gallegati, M. Pareto's law of income distribution: Evidence for Germany, the United Kingdom, and the United States. *Econophysics of wealth distributions: Econophys-Kolkata I*, pp. 3–14, 2005.

Coston, A., Kawakami, A., Zhu, H., Holstein, K., and Heidari, H. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 690–704, 2023. doi: 10.1109/SaTML54575.2023.00050.

Desiere, S. and Struyven, L. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. *Journal of Social Policy*, 50(2):367–385, April 2021. ISSN 0047-2794, 1469-7823. doi: 10.1017/S0047279420000203.

Desiere, S., Langenbucher, K., and Struyven, L. Statistical Profiling in Public Employment Services: An International Comparison. Technical Report 224, OECD Publishing, 2019. URL https://doi.org/10.1787/b5e5f16e-en.

Drezner, Z. and Wesolowsky, G. O. On the Computation of the Bivariate Normal Integral. *Journal of Statistical Computation and Simulation*, 1990.

Elmachtoub, A. N. and Grigas, P. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Fernández-Loría, C. and Provost, F. Causal Decision Making and Causal Effect Estimation Are Not the Same… and Why It Matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022. doi: 10.1287/ijds.2021.0006. URL https://doi.org/10.1287/ijds.2021.0006.

Fischer-Abaigar, U., Kern, C., Barda, N., and Kreuter, F. Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-making in the Public Sector. *Government Information Quarterly*, 41(4):101976, 2024. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2024.101976. URL https://www.sciencedirect.com/science/article/pii/S0740624X24000686.

Guerdan, L., Coston, A., Holstein, K., and Wu, Z. S. Counterfactual Prediction Under Outcome Measurement Error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1584–1598, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594101. URL https://doi.org/10.1145/3593013.3594101.

Jain, S., Creel, K., and Wilson, A. C. Position: Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=44qxX6Ty6F.

Johnson, R. A. and Zhang, S. What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in U.S. Social Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1671–1682, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533223. URL https://doi.org/10.1145/3531146.3533223.

Kallus, N. More Efficient Policy Learning via Optimal Retargeting. *Journal of the American Statistical Association*, 116(534):646–658, 2021. doi: 10.1080/01621459.2020.1788948. URL https://doi.org/10.1080/01621459.2020.1788948.

Kern, C., Bach, R., Mautner, H., and Kreuter, F. When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes. *ACM Journal on Responsible Computing*, 1(4), November 2024. doi: 10.1145/3689485. URL https://doi.org/10.1145/3689485.

Kitagawa, T. and Tetenov, A. Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86 (2):591–616, 2018. doi: https://doi.org/10.3982/ECTA13288. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA13288.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. Prediction Policy Problems. *American Economic Review*, 105(5):491–495, 2015.

Körtner, J. and Bonoli, G. Predictive Algorithms in the Delivery of Public Employment Services. In *Handbook of Labour Market Policy in Advanced Democracies*, pp. 387–398. Edward Elgar Publishing, 2023.

Loxha, A. and Morgandi, M. Profiling the unemployed: a review of OECD experiences and implications for emerging economics. *Social protection discussion papers and notes*, (91051), 2014.

Manski, C. F. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246, 2004. doi: https://doi.org/10.1111/j.1468-0262.2004.00530.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00530.x.

Perdomo, J. C. The Relative Value of Prediction in Algorithmic Decision Making. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Perdomo, J. C., Britton, T., Hardt, M., and Abebe, R. Difficult Lessons on Social Prediction from Wisconsin Public Schools. *arXiv preprint arXiv:2304.06205*, 2023.

Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgenson, E., Mansour, R., and Ghani, R. Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 2039–2047, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2788629. URL https://doi.org/10.1145/2783258.2788629.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A. C., Pentland, A., Polimis, K., Raes, L., Rigobon, D. E., Roberts, C. V., Stanescu, D. M., Suhara, Y., Usmani, A., Wang, E. H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R. B., Argyle, L. P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S. P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I. M., Hummel, L. M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Liu, N., Möser, M., Mack, A. E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K. M., Porter, E., Porter, K. E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J. D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M. K., Woon, W. L., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B. E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B. M., Watts, D. J., and McLanahan, S. Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020. doi: 10.1073/pnas.1915006117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1915006117.

Schmucker, A. and vom Berge, P. Faktisch anonymisierte Version der Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB-Regionalfile) – Version 7521 v1. Forschungsdatenzentrum der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB), 2023a. 10.5164/IAB.SIAB-R7521.de.en.v1.

Schmucker, A. and vom Berge, P. Sample of Integrated Labour Market Biographies Regional File (SIAB-R) 1975 - 2021. FDZ-Datenreport 07/2023 (en), Research Data Centre of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB), Nürnberg, 2023b. 10.5164/IAB.FDZD.2307.en.v1.

Shirali, A., Abebe, R., and Hardt, M. Allocation Requires Prediction Only if Inequality Is Low. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=WUicA0hOF9.

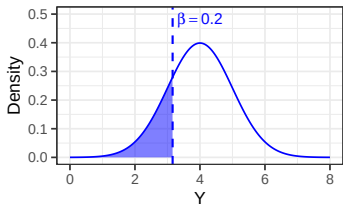
Sun, T. Q. and Medaglia, R. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, 36(2):368–383, 2019.

Wang, A., Kapoor, S., Barocas, S., and Narayanan, A. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsib. Comput.*, 1(1), March 2024. doi: 10.1145/3636509. URL https://doi.org/10.1145/3636509.

Wilder, B. and Welle, P. Learning treatment effects while treating those in need. *arXiv preprint arXiv:2407.07596*, 2024.

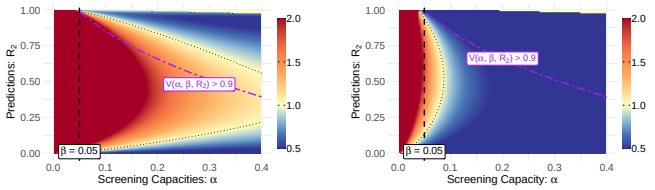
Wirtz, B. W., Weyerer, J. C., and Geyer, C. Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, 42(7):596–615, 2019.

A. Theoretical Investigation



(a) Normal Welfare Distribution

Figure 6. Normal welfare distribution, with vertical lines marking the quantile cutoff ($\beta = 0.2$). The shaded region to the left of the vertical line represents the worst-off segment of the population.



(a) PAR (b) $1/4 \times$ PAR

Figure 7. **Numerical Simulation of the Prediction-Access Ratio (PAR)**, Equation 3, for $\Delta_{R^2} = \Delta_\alpha = 0.01$ and $\beta = 0.05$. Each point represents a screening capacity α (x-axis) and R^2 value (y-axis), with the color bar showing the PAR clipped to the range $[0.5, 2.0]$. The vertical black line marks β, indicating the threshold above which sufficient resources are available to screen everyone under perfect prediction. Dotted black lines represent PAR $= 1$, where improvements in α and R^2 are equally effective. The purple line marks the region in the (α, R^2) space where the policy value $V(\alpha, \beta, R^2)$ exceeds 0.9. Values above 0.9 are located in the upper-right region beyond the purple line.

B. Experiments

B.1. Experimental Setup and Labor Market Data

The dataset is provided via a Scientific Use File by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) (Schmucker & vom Berge, 2023a;b). It is a 2% weakly anonymized random sample of the complete German labor market records from 1975 to 2017 and contains information on 1,827,903 individuals across 62,340,521 observations (Schmucker & vom Berge, 2023b).

We follow the same set of covariates and aggregation procedure for individual unemployment spells as described in Bach et al. (2023), incorporating demographic characteristics, labor market histories, and information about the most recent job. This results in 56 numerical variables and 24 categorical variables, which are one-hot encoded for model training. Figure 8 shows a histogram of individual unemployment durations, which we use as the basis for constructing the outcome variables. The distribution is right-skewed, with a concentration on short durations near zero and a long tail. Such a pattern is commonly observed in other welfare-related outcomes, such as health or income metrics. We define as prediction target the duration of the unemployment period in days Y, capped at 24 months[3]. Differentiating tail values is less important for

[3] In practice, for a fixed β, the problem can also be framed as a classification task (see Appendix B.5).

decision-making, and capping also allows training across years with varying observation windows.

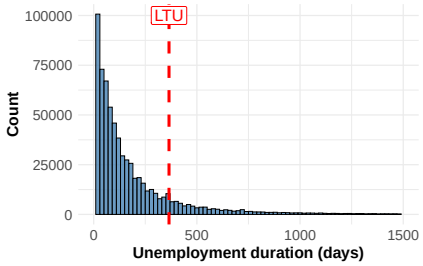


Figure 8. **Unemployment duration** The red line marks the 12 month threshold used to classify a jobseeking episode as long-term unemployment (LTU).

To avoid the impact of significant labor market reforms in Germany and to ensure full observation of unemployment durations up to 24 months, we restrict our analysis to unemployment episodes that began between 2010 and 2015. We use records from 2010 and 2011 to build the training dataset, records from 2012 for validation, and evaluate test performance on data from 2015 (see Figure 9). We left a gap between the training and test data periods to allow enough time for the outcomes in the training data to have been fully observed at test time, in order to mimic a realistic deployment scenario starting at the beginning of 2015.

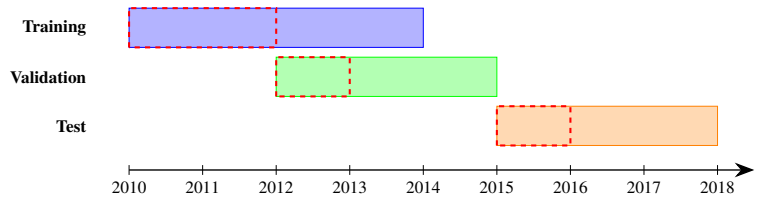


Figure 9. Stacked timeline diagram illustrating training (2010–2013), validation (2012–2014), and test (2015–2017) data periods. Red dashed boundaries within each colored box indicate the possible start dates of unemployment episodes, while the full colored boxes represent the entire observation phases for each dataset.

B.2. Training Details

We use `CatBoost` (https://catboost.ai) for model training. The model was trained for a maximum 5,000 iterations with an early stopping criterion (`early_stopping_rounds` = 20) based on validation performance. Additionally, we train a shallow Decision Tree (`max_depth` = 4) using the `scikit-learn` package. All hyperparameters are kept at their default settings unless otherwise specified.

B.3. Prediction Improvements

To simulate an increase in predictive power by a specified amount Δ_{R^2}, we adjust the model's predictions $\hat{Y}$ using the residuals $Y - \hat{Y}$. Starting with the original predictions $\hat{Y}$ and true outcomes Y, we define the adjusted predictions as

$$\hat{Y}_+ = \hat{Y} + \delta(Y - \hat{Y})$$

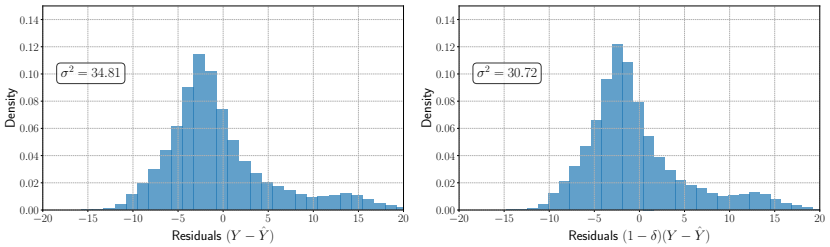
We can then determine the δ corresponding to an increase of Δ_{R^2} in the model's R^2:

$$\delta = 1 - \sqrt{1 - \Delta_{R^2} \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}}$$

For a specified δ, the new residuals are

$$Y - \hat{Y}_+ = (1 - \delta)(Y - \hat{Y})$$

Consequently, the variance decreases by a multiplicative factor: $\mathrm{Var}(Y - \hat{Y}_+) = (1 - \delta)^2 \,\mathrm{Var}(Y - \hat{Y})$.



(a) $\Delta_{R^2} = 0$ (b) $\Delta_{R^2} = 0.1$

Figure 10. **Residual Distribution Before and After Adjustment** Figure (a) shows the residual distribution for the original predictions ($\Delta_{R^2} = 0$), while Figure (b) shows the residual distribution after increasing the R^2-value ($\Delta_{R^2} = 0.1$) for the `CatBoost` model. The adjustment preserves the overall structure of the residuals.

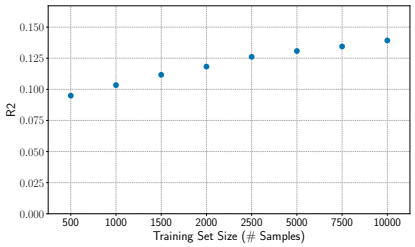
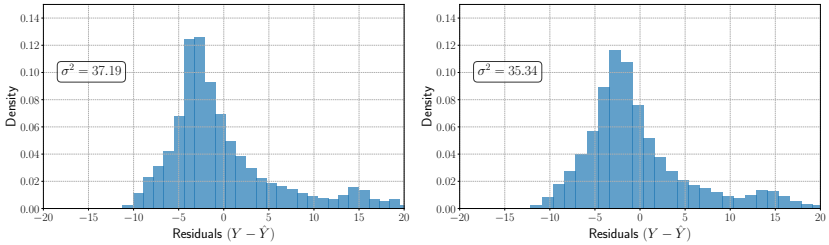


Figure 11. The R^2 value on the test set for varying training set size (`CatBoost` Regression).



(a) 500 Training Samples (b) $10,000$ Training Samples

Figure 12. Residual distributions on the test set for models trained with varying training set sizes.

B.4. Additional Figures

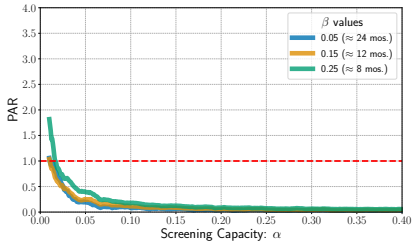


Figure 13. Prediction-Access Ratio for $R^2 = 0$ and $\Delta_{R^2} = \Delta_\alpha = 0.01$.

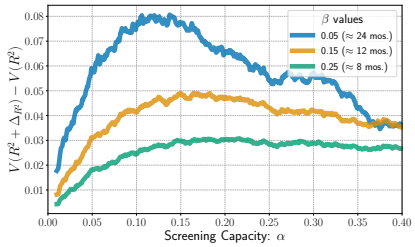


Figure 14. $V(R^2 + \Delta_{R^2}) - V(R^2)$ for `CatBoost` model and $\Delta_{R^2} = 0.05$.

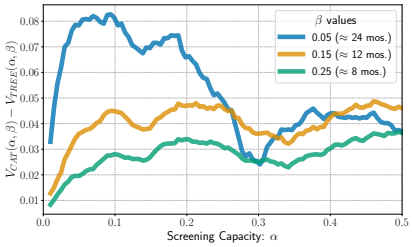
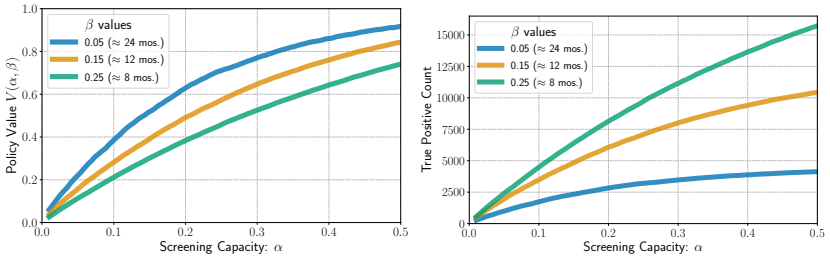


Figure 15. The difference in policy value between a 4-depth decision tree and `CatBoost` model.

B.5. Binary Classification

Instead of predicting the exact duration of unemployment, the problem can be reframed as a binary classification task. For a fixed β, we can define a binary outcome: $Y = 1\{Y \geqslant F^{-1}_{Y,n}(1-\beta)\}$. This approach more directly encodes the target of interest: identifying individuals who may require further screening or assistance. If the chosen classifier provides estimates of class probabilities $\hat{p}(x)$, it can be used to formulate a decision policy $1\{\hat{p}(x) \geqslant F^{-1}_{n,\hat{p}}(1-\alpha)\}$. However, this forces us to specify β and the resulting decision threshold prior to model training. This requirement reduces flexibility compared to a continuous prediction setup, making classification more appropriate when the model is not intended for use in other tasks and when β remains constant across the deployment context. Additionally, directly converting durations to labels discards information on the precise unemployment durations that could be valuable for the modeling process.

As can be seen in Figure 16, the resulting policy values and true positive counts remain very similar compared to the regression case.



(a) Policy Value (b) True Positive Count

Figure 16. Policy Value and True Positive Count on Test Set (Classification).

C. Additional Propositions

Proposition C.1. *(Optimal Policy with Gaussian Error) If $\varepsilon = Y - \hat{Y} \sim \mathcal{N}(0, \gamma^2)$, then the optimal policy $\pi^* : \mathbb{R} \to \{0, 1\}$ to solve the screening problem (Definition 2.1) is equal to:*

$$\pi^*(\hat{Y}_i) = 1\{\hat{Y}_i \leqslant F_{\hat{Y}}^{-1}(\alpha)\}$$

where $F_{\hat{Y}}^{-1}(\alpha)$ is the α-quantile of $\hat{Y}$. The value of the policy is $V(\pi^*) = \Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y \leqslant F_Y^{-1}(\beta)]$.

Proof. Since $Y = \hat{Y} + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \gamma^2)$, it follows for the conditional distribution $Y \mid \hat{Y} \sim \mathcal{N}(\hat{Y}, \gamma^2)$. Since $Y \mid \hat{Y}$ is Gaussian, we can express the conditional probability from Proposition 2.2 in terms of the CDF of the standard normal distribution,

$$\Pr[Y \leqslant F_Y^{-1}(\beta) \mid \hat{Y}] = \Phi\left(\frac{F_Y^{-1}(\beta) - \hat{Y}}{\gamma}\right)$$

To reproduce the ranking induced by $\Pr[Y \leqslant F_Y^{-1}(\beta) \mid \hat{Y}]$, individuals can be ranked in ascending order of $\hat{Y}$. Thus, we can express the optimal policy (Proposition 2.2) in terms of a ranking of $\hat{Y}$,

$$\pi^*(\hat{Y}_i) = 1\{\hat{Y}_i \leqslant F_{\hat{Y}}^{-1}(\alpha)\}$$

where $F_{\hat{Y}}^{-1}(\alpha)$ is the α-quantile of $\hat{Y}$. The value $V(\pi^*)$ that can by achieved by the optimal screening policy π^* can then be expressed as:

$$\begin{aligned} V(\pi^*) = \mathbb{E}\left[\pi^*(\hat{Y}) = 1 \mid Y \leqslant F_Y^{-1}(\beta)\right] &= \mathbb{E}\left[1\{\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha)\} \mid Y \leqslant F_Y^{-1}(\beta)\right] \\ &= \Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y \leqslant F_Y^{-1}(\beta)] \end{aligned}$$

■

D. Proofs

D.1. Optimal Policy for Screening Problem: Proof of Proposition 2.2

Proof. We rewrite the policy value,

$$\begin{aligned} &\mathbb{E}\left[\pi(\hat{Y}_i) = 1 \mid Y \leqslant F_Y^{-1}(\beta)\right] = \tfrac{\mathbb{E}\left[\pi(\hat{Y}_i) 1\{Y \leqslant F_Y^{-1}(\beta)\}\right]}{\Pr[Y \leqslant F_Y^{-1}(\beta)]} \\ &= \tfrac{1}{\beta}\,\mathbb{E}\left[\pi(\hat{Y}_i)\,\mathbb{E}\left[1\{Y \leqslant F_Y^{-1}(\beta)\} \mid \hat{Y}_i\right]\right] \\ &= \tfrac{1}{\beta}\,\mathbb{E}\left[\pi(\hat{Y}_i)\Pr[Y \leqslant F_Y^{-1}(\beta) \mid \hat{Y}_i]\right] \end{aligned}$$

To maximize the objective, individuals $\hat{Y}_i$ with the largest scores $s(\hat{Y}_i) = \Pr[Y \leqslant F_Y^{-1}(\beta) \mid \hat{Y}_i]$ should be prioritized. Thus, the optimal policy is to intervene ($\pi(\hat{Y}_i) = 1$) for the top α-fraction of the population ranked by $\Pr[Y \leqslant F_Y^{-1}(\beta) \mid \hat{Y}]$. ■

D.2. Optimal Policy Value in Gaussian Setting: Proof of Proposition 2.3

Following Proposition C.1, the value of the optimal screening policy π^* can then be expressed as:

$$V(\pi^*) = \Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y \leqslant F_Y^{-1}(\beta)]$$

We can rewrite the conditional probability in terms of the joint distribution of Y and $\hat{Y}$, and note that $\Pr\{Y \leqslant F_Y^{-1}(\beta)\} = F_Y(F_Y^{-1}(\beta)) = \beta$,

$$\Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha) \mid Y \leqslant F_Y^{-1}(\beta)] = \frac{1}{\beta}\Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha), Y \leqslant F_Y^{-1}(\beta)]$$

We then standardize $Y \sim \mathcal{N}(\mu, \eta^2)$ and $\hat{Y} \sim \mathcal{N}(\mu, \eta^2 - \gamma^2)$ and make use that for a normal random variable with mean μ and variance σ^2 the quantile function is $F^{-1}(p) = \mu + \sigma\Phi^{-1}(p)$.

$$\begin{aligned}\frac{1}{\beta}\Pr[\hat{Y} \leqslant F_{\hat{Y}}^{-1}(\alpha), Y \leqslant F_Y^{-1}(\beta)] &= \frac{\Pr\left\{Z_1 \leqslant \frac{F_{\hat{Y}}^{-1}(\alpha)-\mu}{\sqrt{\eta^2-\gamma^2}}, Z_2 \leqslant \frac{F_Y^{-1}(\beta)-\mu}{\eta}\right\}}{\beta} \\ &= \frac{\Pr\left\{Z_1 \leqslant \Phi^{-1}(\alpha), Z_2 \leqslant \Phi^{-1}(\beta)\right\}}{\beta}\end{aligned}$$

Z_1 and Z_2 are standard Gaussian with $\mathrm{Cov}(Z_1, Z_2) = \mathbb{E}[Z_1 Z_2] = \frac{1}{\eta\sqrt{\eta^2-\gamma^2}}\mathrm{Cov}(\hat{Y}, \hat{Y}+\varepsilon) = \frac{\mathrm{Cov}(\hat{Y},\hat{Y})}{\eta\sqrt{\eta^2-\gamma^2}} = \frac{\sqrt{\eta^2-\gamma^2}}{\eta}$. Thus, they are distributed according to a standard bivariate normal distribution with correlation $\rho = \mathrm{Cov}(Z_1, Z_2) = \frac{\sqrt{\eta^2-\gamma^2}}{\eta}$. Thus,

$$V(\pi^*) = \mathbb{E}\left[\pi^*(\hat{Y}) = 1 \mid Y \leqslant F_Y^{-1}(\beta)\right] = \frac{1}{\beta}\Phi_2\left(\Phi^{-1}(\alpha), \Phi^{-1}(\beta); \rho\right)$$

where

$$\Phi_2\left(\Phi^{-1}(\alpha), \Phi^{-1}(\beta)\right) = \int_{-\infty}^{\Phi^{-1}(\alpha)} \int_{-\infty}^{\Phi^{-1}(\beta)} \phi_2(z_1, z_2; \rho)\, \mathrm{d}z_1\, \mathrm{d}z_2$$

and

$$\phi_2(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-1/2(z_1^2 - 2\rho z_1 z_2 + z_2^2)/(1-\rho^2)}$$

D.3. Prediction-Access Ratio for Small Screening Capacities: Proof of Theorem 3.1

Using Taylor's theorem,

$$V(\alpha, \beta, R^2 + \Delta_{R^2}) - V(\alpha, \beta, R^2) = \Delta_{R^2}\frac{\partial}{\partial R^2}V(\alpha, \beta, R^2 + p_{R^2}\Delta_{R^2})$$

where $p_{R^2} \in (0, 1)$. We know from Lemma D.3,

$$\frac{\partial}{\partial R^2}V(\alpha, \beta, R_*^2) \leqslant \frac{1}{\beta\sqrt{8\pi R_*^2(1-R_*^2)}}\phi\left(\frac{\Phi^{-1}(\alpha) - \sqrt{R_*^2}\Phi^{-1}(\beta)}{\sqrt{1-R_*^2}}\right)$$

where $R_*^2 := R^2 + p_{R^2}\Delta_{R^2}$. For $\alpha < 0.5$ and $\beta \leqslant 0.5$, we know $\Phi^{-1}(\alpha) < 0$ and $\Phi^{-1}(\beta) \leqslant 0$. It follows, that for any $\varepsilon_1 > 0$, $0 < R_*^2$ and $0 < \beta$, there exists a threshold value $t_1 > 0$, such that for all $\alpha \leqslant t_1$, we have

$$(1+\varepsilon_1)\frac{\Phi^{-1}(\alpha)}{\sqrt{1-R_*^2}} \leqslant \frac{\Phi^{-1}(\alpha) - \sqrt{R_*^2}\Phi^{-1}(\beta)}{\sqrt{1-R_*^2}} \leqslant (1-\varepsilon_1)\frac{\Phi^{-1}(\alpha)}{\sqrt{1-R_*^2}}$$

If $\alpha < \beta$ we find $\Phi^{-1}(\alpha) - \sqrt{R_*^2}\Phi^{-1}(\beta) < 0$. Since $\phi(x) \leqslant \phi(x')$ for $x \leqslant x' < 0$,

$$\begin{aligned}\frac{1}{\beta\sqrt{8\pi R_*^2(1-R_*^2)}}\phi\left(\frac{\Phi^{-1}(\alpha) - \sqrt{R_*^2}\Phi^{-1}(\beta)}{\sqrt{1-R_*^2}}\right) &\leqslant \frac{1}{\beta\sqrt{8\pi R_*^2(1-R_*^2)}}\phi\left((1-\varepsilon_1)\frac{\Phi^{-1}(\alpha)}{\sqrt{1-R_*^2}}\right) \\ &= \frac{1}{\beta\sqrt{8\pi R_*^2(1-R_*^2)}}\phi\left(\kappa\Phi^{-1}(\alpha)\right) \\ &= \frac{1}{\beta\sqrt{8\pi R_*^2(1-R_*^2)}}\phi\left(\kappa\Phi^{-1}(1-\alpha)\right)\end{aligned}$$

where $\kappa := \frac{(1-\varepsilon_1)}{\sqrt{1-R_*^2}}$. For any $\varepsilon_2 > 0$, there exists a threshold $t_2 > 0$, such that for all $\alpha \leqslant t_2$, we can apply *Lemma B.5.* from Perdomo (2024) to arrive at the following inequality:

$$\phi\left(\kappa\Phi^{-1}(1-\alpha)\right) \leqslant \frac{1}{\sqrt{2\pi}}\left((1+\varepsilon_2)\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\kappa^2}$$

Thus,

$$V(\alpha,\beta,R^2+\Delta_{R^2}) - V(\alpha,\beta,R^2) \leqslant \Delta_{R^2}\frac{1}{\beta 4\pi\sqrt{R_*^2(1-R_*^2)}}\left((1+\varepsilon_2)\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\kappa^2}$$

We can use Taylor's theorem again and from Lemma D.1 we know that

$$\begin{aligned} V(\alpha+\Delta_\alpha,\beta,R^2) - V(\alpha,\beta,R^2) &= \Delta_\alpha\frac{\partial}{\partial\alpha}V(\alpha+p_\alpha\Delta_\alpha,\beta,R^2) \\ &= \Delta_\alpha\frac{1}{\beta}\Phi\left(\frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha+p_\alpha\Delta_\alpha)}{\sqrt{1-R^2}}\right) \end{aligned}$$

where $p_\alpha \in (0,1)$. Since $0<\beta$ and $0<R^2$ there will always be a small enough $\alpha+\Delta_\alpha$ such that

$$\Phi^{-1}(\beta) - \sqrt{R^2}\Phi^{-1}(\alpha+p_\alpha\Delta_\alpha) \geqslant 0$$

Since $\Phi(x) \geqslant 1/2$ for $x \geqslant 0$, it follows

$$\frac{\Delta_\alpha}{2\beta} \leqslant V(\alpha+\Delta_\alpha,\beta,R^2) - V(\alpha,\beta,R^2)$$

It follows for the prediction-access ratio,

$$\frac{\Delta_\alpha}{\Delta_{R^2}}2\pi\sqrt{R_*^2(1-R_*^2)}\left((1+\varepsilon_2)\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{-(1-\varepsilon_1)^2\frac{1}{1-R_*^2}} \leqslant \frac{V(\alpha+\Delta_\alpha,\beta,R^2)-V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta_{R^2})-V(\alpha,\beta,R^2)}$$

For small α, $\Phi^{-1}(1-\alpha)$ grows asymptotically like $\sqrt{\log(1/\alpha)}$. Consequently, the polynomial growth of $\alpha^{-1/(1-R^2)}$ drives the PAR to increase rapidly as α decreases. Since $\frac{1}{1-R_*^2}$ increases with R_*^2 and $R^2 \leqslant R_*^2$, we can lower bound the PAR by inserting R^2 instead of R_*^2:

$$\frac{\Delta_\alpha}{\Delta_{R^2}}2\pi\sqrt{R^2(1-R^2)}\left((1+\varepsilon_2)\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{-(1-\varepsilon_1)^2\frac{1}{1-R^2}} \leqslant \frac{V(\alpha+\Delta_\alpha,\beta,R^2)-V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta_{R^2})-V(\alpha,\beta,R^2)}$$

We can simplify the lower-bound by noting that $0<\varepsilon_1$ and $0<\varepsilon_2$ can be made arbitrarily small by selecting a sufficiently small threshold t for $\alpha+\Delta_\alpha$. Specifically, $\varepsilon_2 < 1$ holds for $\alpha \leqslant 0.05$ (see *Lemma A.6* in Perdomo (2024)).

$$\frac{\Delta_\alpha}{\Delta_{R^2}}\sqrt{R^2(1-R^2)}\left(5.1\cdot\alpha\Phi^{-1}(1-\alpha)\right)^{-\frac{1}{1-R^2}+o(1)} \leqslant \frac{V(\alpha+\Delta_\alpha,\beta,R^2)-V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta_{R^2})-V(\alpha,\beta,R^2)}$$

D.4. Maximally Effective (Local) Prediction Improvements: Proof of Theorem 3.2

We know from Lemma D.2,

$$\begin{aligned} \lim_{\Delta\to 0}\frac{V(\alpha,\beta,R^2+\Delta)-V(\alpha,\beta,R^2)}{\Delta} &= \frac{\partial}{\partial R^2}V(\alpha,\beta,R^2) \\ &= \frac{1}{2\beta\sqrt{R^2}}\phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right) \end{aligned}$$

We insert $\phi_2(\cdot)$ and arrive at

$$\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2) = \underbrace{\frac{1}{4\pi\beta\sqrt{R^2(1-R^2)}}}_{T_1} \times \underbrace{\exp\left(-\frac{1}{2(1-R^2)}(\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta))\right)}_{T_2}$$

The prefactor T_1 diverges as $R^2 \to 1$ or $R^2 \to 0$.

If $R^2 \to 1$, the exponential term will generally suppress the polynomial growth of the prefactor. However for $\alpha = \beta$, we find for the exponent

$$\begin{aligned}-\frac{1}{2(1-R^2)}(\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta)) &= -\frac{1-\sqrt{R^2}}{1-R^2}\Phi^{-1}(\beta)^2 \\ &= -\frac{1}{(1+\sqrt{R^2})}\Phi^{-1}(\alpha)^2 \\ &\overset{R^2\to 1}{=} -\frac{1}{2}\Phi^{-1}(\beta)^2\end{aligned}$$

which is finite. Therefore, $\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)$ becomes unboundedly large if $\alpha = \beta$ and $R^2 \to 1$.

If $R^2 \to 0$, the prefector T_1 diverges again to $+\infty$. The exponent then simplifies to

$$-\frac{1}{2(1-R^2)}(\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta)) = -\frac{1}{2}(\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2)$$

If α and β are not set arbitrarily small or large $\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)$ will diverge. The local PAR (Lemma D.1)

$$\begin{aligned}\lim_{\Delta\to 0}\frac{V(\alpha+\Delta,\beta,R^2)-V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta)-V(\alpha,\beta,R^2)} &= \frac{\frac{\partial}{\partial\alpha}V(\alpha,\beta,R^2)}{\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)} \\ &= \frac{\Phi\left(\frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right)}{\frac{1}{2\sqrt{R^2}}\phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)}\end{aligned}$$

approaches zero in both regimes.

D.5. Prediction-Access Ratio for Local Improvements: Proof of Proposition 3.3

We know

$$\lim_{\Delta\to 0}\frac{V(\alpha+\Delta,\beta,R^2)-V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta)-V(\alpha,\beta,R^2)} = \frac{\frac{\partial}{\partial\alpha}V(\alpha,\beta,R^2)}{\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)}$$

Using Lemma D.1 and Lemma D.3 we find a lower bound for the PAR:

$$\underbrace{\sqrt{8\pi R^2(1-R^2)}}_{T_1}\underbrace{\frac{\Phi\left(\frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right)}{\phi\left(\frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right)}}_{T_2} \leqslant \frac{V(\alpha+\Delta_\alpha,\beta,R^2)-V(\alpha,\beta,R^2)}{V(\alpha,\beta,R^2+\Delta_{R^2})-V(\alpha,\beta,R^2)}$$

We then denote $z := \frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}$ and $T_2 = \frac{\Phi(z)}{\phi(z)}$. We know from Lemma D.4 that $\frac{\Phi(z)}{\phi(z)}$ increases with z. It follows that we need to find the smallest possible z to find a lower bound for T_2. Generally, z decreases with α and increases with β. We treat both cases separately:

1. For $\alpha \leqslant \beta$ we find $\frac{\Phi^{-1}(\beta)(1-\sqrt{R^2})}{\sqrt{1-R^2}} \leqslant z$. Since $\frac{1-\sqrt{R^2}}{\sqrt{1-R^2}}$ decreases with R^2 and $\beta \leqslant 0.5$ we can lower bound the expression by setting $R^2 = 0.15$ and $\beta = 0.03$. Thus, $-1.25 \leqslant z$ and $0.59 \leqslant T_2$. Since $0.15 \leqslant R^2 \leqslant 0.85$ we can lower bound the prefactor $1.79 \leqslant T_1$.
2. For $\alpha \leqslant 0.5$, it follows $\frac{\Phi^{-1}(\beta)}{\sqrt{1-R^2}} \leqslant z$ by setting $\Phi^{-1}(\alpha = 0.5) = 0$. Since $0.15 \leqslant \beta$ and $0.2 \leqslant R^2 \leqslant 0.5$, it follows $0.52 \leqslant T_2$ and $2 \leqslant T_1$

In both cases, we can combine the lower bounds of T_1 and T_2 to find

$$1 \leqslant \frac{V(\alpha + \Delta_\alpha, \beta, R^2) - V(\alpha, \beta, R^2)}{V(\alpha, \beta, R^2 + \Delta_{R^2}) - V(\alpha, \beta, R^2)}$$

D.6. Technical Lemmas

Lemma D.1 (Derivative w.r.t. α)**.**

$$\frac{\partial}{\partial \alpha} V(\alpha, \beta, R^2) = \frac{1}{\beta} \Phi \left(\frac{\Phi^{-1}(\beta) - \sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}} \right) \tag{4}$$

Proof. In the Gaussian setting we find for the policy value (Proposition 2.3),

$$V(\alpha, \beta, R^2) = \frac{\Phi_2\left(\Phi^{-1}(\alpha), \Phi^{-1}(\beta); \rho\right)}{\beta}$$

We first apply Leibniz integral rule,

$$\begin{aligned}
\frac{\partial}{\partial \alpha} V(\alpha, \beta, R^2) &= \frac{\partial}{\partial \alpha} \frac{\Phi_2\left(\Phi^{-1}(\alpha), \Phi^{-1}(\beta); \rho\right)}{\beta} \\
&= \frac{1}{\beta} \int_{-\infty}^{\Phi^{-1}(\beta)} \phi_2\left(z_1, \Phi^{-1}(\alpha); \rho\right) \mathrm{d}z_1 \frac{\partial}{\partial \alpha} \Phi^{-1}(\alpha) \\
&= \frac{1}{\beta \phi\left(\Phi^{-1}(\alpha)\right)} \int_{-\infty}^{\Phi^{-1}(\beta)} \phi_2\left(\Phi^{-1}(\alpha), z_2; \rho\right) \mathrm{d}z_2
\end{aligned}$$

We insert the bivariate density $\phi_2(\cdot)$ and substitute $z_2 - \rho\Phi^{-1}(\alpha) = u\sqrt{1-\rho^2}$

$$\begin{aligned}
&\frac{1}{\beta \phi\left(\Phi^{-1}(\alpha)\right)} \int_{-\infty}^{\Phi^{-1}(\beta)} \phi_2\left(\Phi^{-1}(\alpha), z_2; \rho\right) \mathrm{d}z_2 \\
&= \frac{1}{\beta \phi\left(\Phi^{-1}(\alpha)\right)} \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(\beta)} e^{-1/2(z_2^2 - 2\rho z_2 \Phi^{-1}(\alpha) + \Phi^{-1}(\alpha)^2)/(1-\rho^2)} \mathrm{d}z_2 \\
&= \frac{1}{2\pi\beta \phi\left(\Phi^{-1}(\alpha)\right)} \int_{-\infty}^{(\Phi^{-1}(\beta) - \rho\Phi^{-1}(\alpha))/\sqrt{1-\rho^2}} e^{-1/2(u^2(1-\rho^2) + \rho^2\Phi^{-1}(\alpha)^2 + (1-\rho^2)\Phi^{-1}(\alpha)^2)/(1-\rho^2)} \mathrm{d}u \\
&= \frac{1}{2\pi\beta \phi\left(\Phi^{-1}(\alpha)\right)} e^{-1/2\Phi^{-1}(\alpha)^2} \int_{-\infty}^{(\Phi^{-1}(\beta) - \rho\Phi^{-1}(\alpha))/\sqrt{1-\rho^2}} e^{-1/2u^2} \mathrm{d}u \\
&= \frac{1}{\beta} \Phi \left(\frac{\Phi^{-1}(\beta) - \rho\Phi^{-1}(\alpha)}{\sqrt{1-\rho^2}} \right) = \frac{1}{\beta} \Phi \left(\frac{\Phi^{-1}(\beta) - \sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}} \right)
\end{aligned}$$

■

Lemma D.2 (Derivative w.r.t. R^2)**.**

$$\frac{\partial}{\partial R^2} V(\alpha, \beta, R^2) = \frac{1}{2\beta\sqrt{R^2}} \phi_2\left(\Phi^{-1}(\alpha), \Phi^{-1}(\beta)\right) \tag{5}$$

where $\phi_2(\cdot)$ is the standard bivariate density.

Proof.

$$\begin{aligned}\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2) &= \frac{\partial}{\partial R^2}\frac{\Phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)}{\beta}\\ &= \frac{1}{\beta}\frac{\partial\rho}{\partial R^2}\frac{\partial}{\partial\rho}\Phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)\\ &= \frac{1}{2\beta\sqrt{R^2}}\frac{\partial}{\partial\rho}\Phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)\\ &= \frac{1}{2\beta\sqrt{R^2}}\phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)\end{aligned}$$

where $\phi_2(\cdot)$ is the standard bivariate density. We utilized $R^2=\rho^2$, and in the final step applied the partial derivative of the standard bivariate cumulative distribution with respect to its correlation ρ (Drezner & Wesolowsky, 1990). ■

Lemma D.3 (Upper bound of R^2 derivative)**.** *Let* $0\leqslant\sqrt{R^2}\leqslant 1$*. Then,*

$$\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)\leqslant\frac{1}{\beta\sqrt{8\pi R^2(1-R^2)}}\phi\left(\frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right) \tag{6}$$

$$\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)\leqslant\frac{1}{\beta\sqrt{8\pi R^2(1-R^2)}}\phi\left(\frac{\sqrt{R^2}\Phi^{-1}(\beta)-\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right) \tag{7}$$

Proof. We know from Lemma D.2,

$$\begin{aligned}\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2) &= \frac{1}{2\beta\sqrt{R^2}}\phi_2\left(\Phi^{-1}(\alpha),\Phi^{-1}(\beta);\rho\right)\\ &= \frac{1}{4\pi\beta\sqrt{R^2(1-R^2)}}\\ &\quad\times\exp\left(-\frac{1}{2(1-R^2)}(\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta))\right)\end{aligned}$$

Since $0\leqslant\sqrt{R^2}\leqslant 1$,

$$\begin{aligned}\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta) &\geqslant R^2\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta)\\ &= (\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha))^2\geqslant 0\end{aligned}$$

Similarly,

$$\Phi^{-1}(\alpha)^2+\Phi^{-1}(\beta)^2-2\sqrt{R^2}\Phi^{-1}(\alpha)\Phi^{-1}(\beta)\geqslant(\sqrt{R^2}\Phi^{-1}(\beta)-\Phi^{-1}(\alpha))^2\geqslant 0$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2) &\leqslant \frac{1}{4\pi\beta\sqrt{R^2(1-R^2)}}\exp\left(-\frac{1}{2(1-R^2)}(\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha))^2\right)\\ &= \frac{1}{\beta\sqrt{8\pi R^2(1-R^2)}}\phi\left(\frac{\Phi^{-1}(\beta)-\sqrt{R^2}\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right)\end{aligned}$$

and

$$\frac{\partial}{\partial R^2}V(\alpha,\beta,R^2)\leqslant\frac{1}{\beta\sqrt{8\pi R^2(1-R^2)}}\phi\left(\frac{\sqrt{R^2}\Phi^{-1}(\beta)-\Phi^{-1}(\alpha)}{\sqrt{1-R^2}}\right)$$

■

Lemma D.4. *The ratio*

$$\frac{\Phi(z)}{\phi(z)} \tag{8}$$

is increasing in z*.*

Proof. We compute the derivative of the ratio $\frac{\Phi(z)}{\phi(z)}$,

$$\frac{\partial}{\partial z}\frac{\Phi(z)}{\phi(z)} = \frac{\phi^2(z) + z\phi(z)\Phi(z)}{\phi^2(z)} = 1 + z\frac{\Phi(z)}{\phi(z)}$$

For $z \geqslant 0$ the derivative is clearly positive. For $z < 0$ we start by rewriting,

$$1 + z\frac{\Phi(z)}{\phi(z)} = \frac{z}{\phi(z)}\left(\frac{\phi(z)}{z} + \Phi(z)\right)$$

Since $\frac{\partial}{\partial z}\left(\frac{\phi(z)}{z} + \Phi(z)\right) = \frac{-z^2\phi(z) - \phi(z)}{z^2} + \phi(z) = \frac{-\phi(z)}{z^2} < 0$ and $\frac{\phi(z)}{z} + \Phi(z) \overset{z \to -\infty}{\to} 0$, it follows for $z < 0$ that $\left(\frac{\phi(z)}{z} + \Phi(z)\right) < 0$. Thus for any z,

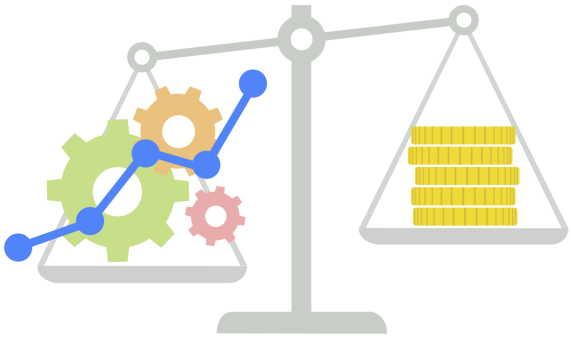
$$\frac{\partial}{\partial z}\frac{\Phi(z)}{\phi(z)} \geqslant 0$$

■

7 On the Meta-Design of Allocation Problems

Publication

U. Fischer-Abaigar, E. Aiken, C. Kern, and J. C. Perdomo (2026). "On the Meta-Design of Allocation Problems". *arXiv preprint arXiv:2602.08786*. arXiv: 2602.08786



Software rvp[1] is a small Python toolkit for studying the meta-design of allocation problems. It helps evaluate how downstream welfare changes when a planner varies design choices such as capacity constraints, available data, prediction, and treatment quality.

Author Contributions UFA led the project as first author, ran all experiments on the labor market data, conceptualized and developed the accompanying software toolkit, integrated the poverty targeting experiments into the toolkit, wrote the initial draft of all sections except the section on poverty targeting, and created all visualizations. EA conducted the initial poverty targeting experiments, wrote the initial draft of the poverty targeting section, and contributed to conceptualization, reviewing, and editing throughout. JCP contributed to the initial conceptualization, supervised the project, and provided feedback and editing throughout. CK contributed to conceptualization, reviewing, and editing.

[1] https://github.com/unai-fa/relative-value-of-prediction

On the Meta-Design of Allocation Problems

Unai Fischer-Abaigar*[1,2], Emily Aiken[3], Christoph Kern[1,2], and Juan Carlos Perdomo[4, 5]

[1]*LMU Munich*
[2]*Munich Center for Machine Learning*
[3]*University of California San Diego*
[4]*New York University*
[5]*Massachusetts Institute of Technology*

arXiv:2602.08786v4 [cs.CY] 8 May 2026

Abstract

There is an extensive literature that studies how to find optimal policies in resource allocation problems, taking the underlying design parameters that define the allocation, such as what data is collected, how many people can be served, and quality of service as fixed constraints. Yet, from a planner's perspective, these design parameters are themselves optimization variables that are just as important in determining overall welfare as selecting the optimal targeting rule for a given set of constraints. This realization motivates a rich set of meta-design questions exploring how planners should make principled decisions about investments in prediction, capacity constraints, and treatment quality, all of which lie upstream of classical policy optimization. Building on initial theoretical work in this space, our paper has three main contributions. First, we formally define the broad meta-design space of resource allocation problems. Second, we develop empirical tools that enable practitioners to reliably navigate it. Third, we demonstrate the framework in two real-world case studies on German employment services and targeted cash transfer programs in Ethiopia.

1 Introduction

Public institutions increasingly use predictive algorithms to guide the allocation of scarce societal resources. For instance, algorithms are used to target tutoring services to students at risk of dropout [Perdomo et al., 2025], allocate public housing units [Vayanos et al., 2026], and to prioritize vulnerable communities for early access to new vaccines [Barda et al., 2020]. Given a specification for an allocation problem — that is, a set of data measurements (features, outcomes), utility function, and limit on the number of individuals intervened on — there is a large body of work that studies how to learn optimal decision rules that specify who in the population should receive resources. These studies range from learning optimal treatment assignment rules through empirical welfare maximization [Athey and Wager, 2021, Kitagawa and Tetenov, 2018], to training constrained predictors that satisfy operational or fairness requirements [Cotter et al., 2019], to establishing calibration conditions under which post-processing of predictions leads to optimal downstream decisions [Hu et al., 2023].

*Corresponding author: Unai.FischerAbaigar@stat.uni-muenchen.de

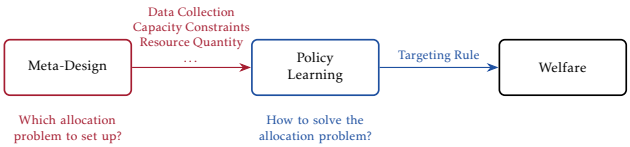


Figure 1: A planner shapes the allocation problem through investment decisions, which feed into a policy learning algorithm that produces a targeting rule and determines the resulting welfare.

Before deciding on a targeting rule, a social planner inevitably needs to first decide on the parameters that define the allocation (see Figure 1). What data is collected, how many people can be served, and the quality of interventions are ultimately design choices of a social planner. Furthermore, these *meta-design* decisions are just as if not more important than the downstream choice of decision rule. For instance, a planner who decides to expand program capacity while keeping the targeting rule fixed may do far more good than one who perfects the targeting rule for a program that is under-allocating.

Our Contributions. Given this background, our work has three main contributions. First, we formally define the space of questions in the meta-design of data-driven allocation problems. Second, we provide tools that enable practitioners to reliably understand which components of their allocation system matter most for downstream welfare, and how to prioritize investments across them. And lastly, we illustrate our methodology in the context of two real-world case studies on German employment services and poverty targeting in Ethiopia.

On a technical level, our analysis centers on an object we term the *welfare surface*. Intuitively, the welfare surface describes how the value achieved by the optimal targeting rule for a given choice of design parameters varies with these meta-design parameters (capacity constraints, data distributions, treatment effects). Drawing on connections to classical microeconomic theory and nonlinear optimization, we show how the welfare surface empowers planners to address a range of meta-design questions. These include which design axes matter most in their current configuration, testing whether existing design choices are optimal given fixed budgets, and how to allocate budgets among competing investments when expanding a system.

As a final note, a key insight of our work is that meta-design conclusions are inherently context-dependent. As shown via our case studies, which investments matter most depends on a planner's current configuration, data, and constraints in ways that no general theory can fully anticipate. Rather than prescribing one-size-fits-all answers, we give planners the tools to derive rigorous conclusions that are valid in their own context and the framework to know when those conclusions change. Importantly, we demonstrate this on rich real-world data drawn directly from the institutional settings we study, from household poverty surveys used to target cash transfers to real-world governmental labor market records used to profile job seekers.

2 Related Work

Prediction-Access Ratio. The most related prior investigations are Perdomo [2024] and Fischer-Abaigar et al. [2025]. They introduce the prediction-access ratio (PAR) to study the relative welfare gains achieved by improving prediction and expanding program capacity in resource

allocation contexts. Their main results identify conditions under which the relative value of prediction is low in stylized theoretical models. We build on this work in two directions. First, we explore a broader space of meta-design questions, not just the relative welfare gains between prediction and access. Second, their analysis primarily relies on stylized Gaussian models, which allow for sharp theoretical insights but constrain external validity. We show how a planner can explore the design space of their own program, without relying on whether its primitives happen to match a prior theoretical model.

Design Space of Allocations. A wider literature points to the size of this design space by examining different components of allocation systems. Perdomo et al. [2025] find that early warning systems in Wisconsin schools can deliver much of their value without individual-level risk prediction, while Mashiat et al. [2025] find that individually targeted eviction-prevention outreach substantially outperforms neighborhood-based targeting. On the theoretical side, Shirali et al. [2024] show that individual targeting only outperforms aggregate targeting in high-budget regimes when inequality is low, and Casacuberta and Hardt [2026] find that learning a good allocation can require less data than estimating heterogeneous effects. Other contributions examine alternative notions of welfare, including frameworks that motivate randomization in allocation [Jain et al., 2024].

Solving Allocation Problems. A related literature focuses on *solving* allocation problems once formulated. Common approaches include learning risk scores to rank individuals [Kleinberg et al., 2015, Liu et al., 2026, Wang et al., 2024], estimating counterfactual risk or treatment effects to account for causal structure [Coston et al., 2020, Wager and Athey, 2018, Künzel et al., 2019], and learning optimal treatment rules directly [Kitagawa and Tetenov, 2018, Athey and Wager, 2021, Manski, 2004, Kallus, 2018]. Decision-focused learning trains predictors to account for downstream optimization [Elmachtoub and Grigas, 2022, Ren et al., 2024]. Recent work clarifies how these perspectives relate, connecting estimation targets to utility under different problem formulations [Sharma and Wilder, 2025, Liu et al., 2026, Fischer-Abaigar et al., 2024, Kern et al., 2025].

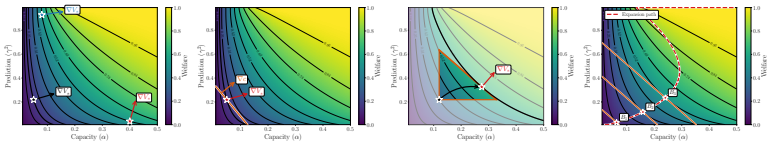
Our work is not primarily concerned with solving a given allocation problem, but rather with the meta-design of the allocation problem itself and the tradeoffs that arise from different design choices.

3 The Meta-Design of Resource Allocation Problems

Here, we formalize the meta-design space of resource allocation problems, introduce the welfare surface as the central analytical object, and outline the questions a planner can answer by studying it. We start our presentation by defining the optimal social welfare achieved under a fixed set of design parameters. Throughout our work, we focus on allocation problems where a social planner assigns a single scarce resource to individuals in a population, as defined below:

Definition 3.1 (Allocation Problem)**.** *Let $\mathcal{D}$ be a distribution over pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and let $\alpha \in (0, 1)$ denote a capacity constraint. Fix a utility function $u : \mathcal{Y} \times \{0, 1\} \to \mathbb{R}$. We define $V_*(\mathcal{D}, u, \alpha)$ to be the value of the optimal policy $\pi : \mathcal{X} \to [0, 1]$ with respect to $(\mathcal{D}, u, \alpha)$,*

$$V_*(\mathcal{D}, u, \alpha) := \max_{\pi:\, \mathcal{X} \to [0,1]} \mathbb{E}_{\mathcal{D}}[u(y, \pi(x))] \quad s.t.\ \mathbb{E}_{\mathcal{D}}[\pi(x)] \leqslant \alpha$$



(a) Local Rate of Improvement (b) Inspecting Optimality (c) Budget Optimization (d) Expansion Path

Figure 2: Welfare surface for the stylized Gaussian model of Fischer-Abaigar et al. [2025], over program capacity α and prediction quality (γ^2 coefficient of determination). Black lines mark iso-welfare contours. **(a)** Normalized welfare gradients at three configurations. **(b)** When the welfare gradient is not aligned with the cost gradient (orange line, constant cost), the configuration is not locally optimal. **(c)** Feasible region (orange) under additive linear costs from an existing system (black star). **(d)** The red line traces the optimal configuration as the budget grows.

The allocation problem is defined based on three primitives; the utility function u, a distribution $\mathcal{D}$ which implicitly defines the set of features and outcomes, and the capacity constraint α. The outcome y and the utility function u together encode the social planner's objective. Given these three primitives, $V_*(\mathcal{D}, u, \alpha)$ denotes welfare under the optimal targeting rule.

Example 3.2 (Targeting the Worst-Off)**.** Consider the problem of allocating attention to individuals deemed to be "worst-off" according to outcomes y, e.g., job seekers at risk of long-term unemployment. There, $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathbb{R}$ and the utility function $u(y, a)$ is equal to $a \cdot \mathbf{1}\{y \geqslant t_\beta\}$ where t_β is a pre-specified cutoff and $a \in \{0, 1\}$. Learning plays a role in these settings since the optimal targeting rule is to first approximate $f_*(x) = \Pr[y \geqslant t_\beta | x]$, and then allocate to individuals with the highest values of $f_*(x)$. That is, $\pi_*(x) = \mathbf{1}\{f_*(x) \geqslant t_\alpha\}$ where t_α is chosen to satisfy the capacity constraint in Definition 3.1.

In practice, π_* is unknown and must be learned from finite-samples. As we discussed, there is by now a rich literature [Bhattacharya and Dupas, 2012, Slivkins et al., 2023, Hu et al., 2023] that given as input $(\mathcal{D}, u, \alpha)$ proposes algorithms for learning π_* using i.i.d samples drawn from $\mathcal{D}$. In our work, we make use of these algorithms as a subroutine and focus on the task of optimizing over the choice of parameters $(\mathcal{D}, u, \alpha)$ themselves.

The Meta-Design Universe. When building a resource allocation system, the first order of business for a planner is to decide on the resources to provide to selected individuals (for instance, the dollar amount in cash transfers), the goals of the system (jointly encoded in a utility function u), the information available for targeting (a data distribution $\mathcal{D}$), and the scale of the program (capacity constraints α). In other words, planners start by selecting a precise choice of design parameters $(\mathcal{D}, u, \alpha)$ from a set, or *design universe*, $\mathcal{U}$ of parameters.

The structure of the set $\mathcal{U}$ depends on the planner's situation. When improving an existing system, $\mathcal{U}$ is naturally defined as the set of configurations reachable from a current baseline $(\mathcal{D}_0, \alpha_0, u_0)$ via available *policy levers*, each lever defining an axis of improvement along which the planner can invest more or less. For example, expanding capacity by $\Delta_\alpha \in [0, \bar{\alpha}]$ traces out a path $\mathcal{U}_{\text{cap}} = \{(\mathcal{D}_0, u_0, \alpha_0 + \Delta_\alpha) : \Delta_\alpha \in [0, \bar{\alpha}]\}$. Potential improvements in prediction define a set $\mathcal{U}_{\text{pred}} = \{(\mathcal{D}_\lambda, u_0, \alpha_0,) : \lambda \in \Lambda\}$, where $\mathcal{D}_\lambda$ is a family of distributions. For example, λ could index

distributions over different feature sets. When considering improvements along multiple axis, the combined design universe is the joint parameter space, for example $\mathcal{U} = \{(\mathcal{D}_\lambda, \alpha_0, u_0) : \lambda \in \Lambda,\ \Delta_\alpha \in [0, \bar{\alpha}]\}$.

Whether building a system from scratch or improving an existing system, the design universe $\mathcal{U}$ will typically be subject to a total budget B, $\mathcal{U}_B = \{(\mathcal{D}, u, \alpha) \in \mathcal{U} : c(\mathcal{D}, u, \alpha) \leqslant B\}$ where $c(\mathcal{D}, u, \alpha)$ is the implementation cost of a given configuration. The planner's meta-design problem is then to select an element of $\mathcal{U}_B$ that maximizes downstream welfare $V_*(\mathcal{D}, u, \alpha)$.

The Welfare Surface. Each element of the design universe $\mathcal{U}$ corresponds to a distinct allocation problem, and therefore a distinct level of downstream welfare under the optimal targeting rule. This mapping $(\mathcal{D}, \alpha, u) \to V_*(\mathcal{D}, \alpha, u)$ for $(\mathcal{D}, u, \alpha) \in \mathcal{U}$ defines the *welfare surface* describing the performance of the allocation system under different configurations. Intuitively, the welfare surface is a landscape over the space of design parameters: high regions correspond to configurations that achieve strong downstream welfare. And as we will illustrate next, the geometry of the surface, such as its gradients and contour lines, provides insight into the relative value of different policy levers.

Given a dataset $S = \{(x_i, y_i)\}_{i=1}^n$ the welfare surface can be *empirically computed* for practically relevant design universes. For example, one can *a*) simulate welfare under expanded access $\alpha \to \alpha + \Delta$, allocating to more individuals, and *b*) simulate the effects of enhanced interventions (*e.g.* model the impact of a larger cash transfer under a specific choice of utility), or even *c*) training under different distributions of feature sets. In Section 4.1, we compute welfare surfaces over data coverage, share of households eligible, and cash transfer size for poverty targeting in Ethiopia. In Section 4.2, we do the same over institutional capacity and feature collection for the profiling of unemployed job seekers in Germany. As part of our work, we also provide open source software[1] than enables practitioners to compute the design surface for common sets $\mathcal{U}$.

This stands in contrast to earlier theoretical work, which analyzed the relative value of policy levers under stylized assumptions such as Gaussian data and linear utilities [Perdomo, 2024, Fischer-Abaigar et al., 2024]. Our approach lets planners ask a broader set of questions directly in their own problem settings, without first checking whether their data matches these prior stylized settings.

3.1 Navigating the Meta-Design Universe

Given a design space of feasible allocation problems $\mathcal{U}$, there are a number of questions that a social planner needs to answer. These include:

(Q1) Given my current system, what kinds of local improvements are most effective?

(Q2) Is my current configuration optimal for my budget?

(Q3) How should I choose parameters optimally?

Given the framework we have set up so far, we now show how these questions can be answered by inspecting the welfare surface. As a means of presentation, we illustrate an overview of this methodology in the context of the theoretical model presented in Fischer-Abaigar et al. [2024], showing how it generalizes their analysis. They consider allocation systems that aim to identify individuals in need (see Example 3.2), and model outcomes y and predictions $\hat{y}$ as

[1] Available at https://github.com/unai-fa/relative-value-of-prediction.

following a bivariate normal distribution $(y, \hat{y}) \sim \mathcal{D}_\gamma$ with correlation γ. The design axes they analyze are the capacity of the allocation system α and stylized improvements in prediction by varying the squared-correlation (or r^2 of the predictions) γ^2, implying a design universe $\mathcal{U} = \{\mathcal{D}_\gamma, u, \alpha : \gamma \in [0,1],\ \alpha \in [0,1]\}$.

More generally, when the design universe admits a suitable parametrization, we can index each design configuration $(\mathcal{D}, u, \alpha)$ by a vector $\theta \in \Theta = \Theta_1 \times \cdots \times \Theta_d \subseteq \mathbb{R}^d_{\geqslant 0}$ where each coordinate $\theta_i \in \Theta_i \subseteq \mathbb{R}_{\geqslant 0}$ corresponds to one design axis. In Fischer-Abaigar et al. [2025], $\theta = (\alpha, \gamma^2) \in [0,1]^2$ and $V_*(\theta) := V_*(\alpha, \gamma^2) = V_*(\mathcal{D}_\gamma, u, \alpha)$ for a fixed utility $u(y, a) = a \cdot \mathbf{1}\{y \geqslant t_\beta\}$. Figure 2 shows the corresponding welfare surface. In Section 4, we will instantiate our framework empirically, computing welfare surfaces from real data and exploring more realistic improvement axes.

Local Improvements (Q1). Given a parameterization θ of the design universe $\mathcal{U}$, the gradients of the welfare surface encode the magnitudes of incremental improvements along various axes,

$$\nabla V_*(\alpha, \gamma^2) = \left(\frac{\partial}{\partial \alpha} V_*(\alpha, \gamma^2), \frac{\partial}{\partial \gamma^2} V_*(\alpha, \gamma^2) \right)^\top$$

The two components capture the marginal value of expanding access and the marginal value of improving prediction, respectively.

As an analytical tool, the gradient of the welfare surface enables planners to directly visualize, and communicate to others, which design parameters are most important at a particular configuration. In particular, as seen in Figure 2(a), the gradient of the welfare surface evaluated at a particular point is perpendicular to the contour line of that point. Vertical gradients, such as those on the bottom right of the plot indicate that a small improvement in prediction are relatively much more valuable than small improvements in access. Horizontal gradients, such as those on the left side, indicate the opposite: the marginal benefit of expanding access far outweighs that of expanding prediction.

Our framing generalizes the insights from prior theoretical work. In particular, Perdomo [2024], Fischer-Abaigar et al. [2024] formalize the relative value of prediction in allocation problems via a notion they term the *prediction-access ratio* (PAR), defined as,

$$\mathsf{PAR}(\alpha, \gamma^2) = \frac{\partial}{\partial \alpha} V_*(\alpha, \gamma^2) \,/\, \frac{\partial}{\partial \gamma^2} V_*(\alpha, \gamma^2)$$

This is exactly equal to the ratio of the entries in the gradient of the welfare surface for this specific choice of a design universe $\mathcal{U}$. Their main results consist of showing that this ratio is very large for small values of α: an insight immediately visible from the near-flat gradients on the left of Figure 2(a), which imply that the marginal welfare value of expanding capacity far exceeds that of improving the predictor.

Relative to prior work, our framework extends this analysis in two ways: it applies to any pair of design axes rather than just prediction versus capacity, and it can be computed empirically on data directly corresponding to a user's domain rather than requiring modeling assumptions. In particular, the ratio of any two partial derivatives of V_* defines a *marginal rate of substitution*,

$$\mathsf{MRS}_{ij}(\theta) = \frac{\partial}{\partial \theta_i} V_*(\theta) \,/\, \frac{\partial}{\partial \theta_j} V_*(\theta)$$

In microeconomic theory, the MRS between i and j denotes the rate at which a consumer can exchange good i for good j without changing welfare [Mas-Colell et al., 1995]. In allocations,

the MRS between two policy levers (of which the PAR is just a special case) specifies the rate at which a planner is willing to tradeoff two design parameters (e.g. prediction versus access) to maintain the same performance. As we will now show, the MRS gives practioners insight into whether their systems are cost-optimal.

Inspecting Optimality (Q2). When a planner has a fixed budget B and cost function $c : \Theta \to \mathbb{R}_{\geqslant 0}$, mapping design parameters to real numbers, the welfare surface enables planners to test whether their local configurations are suboptimal. In particular, consider the problem of checking whether a particular parameter vector $\theta \in \Theta = \Theta_1 \times \cdots \times \Theta_d$ is optimal for a given budget constraint. That is, we want to test the following: $\theta^* \in \arg\max_{\theta \in \Theta,\, c(\theta) \leqslant B} V_*(\theta)$.

Classical nonlinear optimization theory states that any solution θ^* to this problem must satisfy the following first-order conditions, whenever the KKT conditions apply:

Fact 3.3. *Suppose $\Theta = [0,1]^d$, and let $V_* : \Theta \to \mathbb{R}$ and $c : \Theta \to \mathbb{R}_{\geqslant 0}$ be continuously differentiable. Suppose that θ^* solves*

$$\max_{\theta \in \Theta,\, c(\theta) \leqslant B} V_*(\theta)$$

and suppose that a standard constraint qualification holds at θ^*, as is the case here when c is linear in θ. Then there exists a budget multiplier $\lambda \geqslant 0$, with $\lambda(c(\theta^*) - B) = 0$, such that, for each coordinate $i \in \{1, \ldots, d\}$, $\frac{\partial V_*}{\partial \theta_i}(\theta^*) = \lambda \frac{\partial c}{\partial \theta_i}(\theta^*)$ if $\theta_i^* \in (0,1)$, $\frac{\partial V_*}{\partial \theta_i}(\theta^*) \geqslant \lambda \frac{\partial c}{\partial \theta_i}(\theta^*)$ if $\theta_i^* = 1$, and $\frac{\partial V_*}{\partial \theta_i}(\theta^*) \leqslant \lambda \frac{\partial c}{\partial \theta_i}(\theta^*)$ if $\theta_i^* = 0$.

An immediate implication is that, at an optimal design, the marginal rate of substitution between any two design axes that are not at their bounds must equal the ratio of their marginal costs.

Corollary 3.4. *Let $\theta^* \in \Theta$ satisfy the conditions of Fact 3.3. Then for any pair of coordinates (i, j) with $0 < \theta_i^* < 1$ and $0 < \theta_j^* < 1$ for which the MRS and the cost ratio are well-defined, we have*

$$\mathsf{MRS}_{ij}(\theta^*) = \frac{\partial}{\partial \theta_i} c(\theta^*) \,/\, \frac{\partial}{\partial \theta_j} c(\theta^*)$$

Figure 2(b) provides a visual illustration of this condition. Intuitively, if the marginal rate of substitution between two design axes is not equal to the marginal ratio of costs, then the planner is better off by reducing their investment along one dimension and reallocating it to another. For instance, in Figure 2(b), the configuration at the top of the feasible region (blue) is locally inefficient; the planner is better off investing less in prediction and expanding access. Doing so would not increase their costs, but significantly improve the overall performance of their system.

We remark that using our software toolkit, planners can input their own cost functions and check whether these optimality conditions hold in the context of their own planning problems. We illustrate this methodology in the context of various case studies in the next section.

Choosing Optimal Design Parameters (Q3). A key advantage of the welfare surface framework is that it cleanly separates two concerns: understanding how welfare varies as a function of design choices, and understanding what those choices cost. The welfare surface $V_*(\theta)$ is a property of the allocation problem alone and can be computed by practitioners on a fixed dataset

and over a wide range of design universes $\mathcal{U}$, independently of any cost considerations. Budget constraints enter only as a second step, as a feasible region $\mathcal{U}_B = \{\theta \in \mathcal{U} : c(\theta) \leqslant B\}$ overlaid on top of the surface (see Figure 2(c)). Finding the optimal design then reduces to identifying the point in $\mathcal{U}_B$ that lies on the highest iso-welfare contour, a problem addressed by our software and which is efficiently solvable due to the often low-dimensional structure of $\mathcal{U}$.

This separation is practically valuable. Beyond testing for local optimality as in Q2, a planner can compute the welfare surface once and then evaluate it under many different cost structures or budget levels. As the budget B grows, the feasible region expands and the optimal configuration traces an *expansion path* through $\mathcal{U}$ (see Figure 2(d)). The shape of this path reveals how the optimal mix of investments shifts with scale — for instance, whether data collection should precede capacity expansion at low budgets, or whether the two should be pursued in parallel.

4 Empirical Case Studies

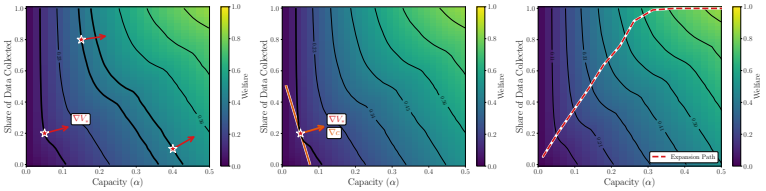
As demonstrated, the geometry of the welfare surface encodes answers to a range of design questions that prior analytical work has only addressed in stylized settings. We now turn to practice and compute the welfare surface on real-world data in two case studies, illustrating the kinds of conclusions a planner can draw and how those conclusions shift as the problem specification changes.

4.1 Poverty Targeting in Ethiopia

Direct cash transfer programs are globally ubiquitous, reaching 1.36 billion people across 962 programs in 2020 [Gentilini, 2022]. Many are targeted using proxy means tests (PMTs) [Barrientos, 2018], which use a simple predictive model to infer household poverty and prioritize the poorest of those for cash transfers [Grosh et al., 1995, Brown et al., 2018]. While PMT accuracy has been assessed in several contexts [Brown et al., 2018, Schnitzer and Stoeffler, 2024, McBride and Nichols, 2018], the trade-offs between investing in prediction and allocation in this setting have not previously been analyzed. See Appendix B for additional background.

Allocation Problem. A social planner wants to distribute cash transfers to poor households but cannot directly observe household poverty. Instead, short "poverty scorecard" surveys are collected in the field and a predictive model is trained to estimate household poverty from the survey responses. Households with the highest estimated poverty are then eligible for transfers up to a capacity threshold, $\pi(x) = \mathbf{1}\{f(x) \leqslant t_\alpha\}$. We simulate this approach using Ethiopia's 2015 Living Standards Measurement survey, covering 4,694 households, which contains both poverty scorecard questions (the features x) and measure of household consumption (serving as ground-truth poverty y).

Design Space. Because the scorecard surveys used to estimate household poverty in PMTs are expensive to conduct — with a median per-survey cost of $13 USD PPP reported in the literature [Aiken et al., 2023] — such PMT registry databases often fail to cover all households in a country, or even all households in the poorest parts of a country. A recent review of PMT registries in low- and middle-income countries found a median registry coverage of 21% of households (ranging from less than 1% in Belize to over 99% in Argentina) [Grosh et al., 2022]. Households without



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 3: **Case Study 1.** Welfare surface for a hypothetical poverty targeting program in Ethiopia across the share of data collected and the capacity of the program. Welfare measured as the number of correctly identified individuals below the poverty line (33%). Averaged over 100 cross validation splits. Data collection refers to scorecard surveys at a median cost of $13 USD PPP per household; capacity investment expands the share of households receiving transfers of $100 USD PPP.

data cannot be individually targeted, effectively dramatically eroding accuracy of the targeting model at deployment time[2]. The administrators of PMT-targeted cash transfer programs thus face a trade-off in where to invest resources: they can allocate resources to cash transfers — serving a larger share of households or increasing the amount each recipient receives — or they can expand data collection (that is, measure features for more people), ensuring that more households have data available for prediction.

4.1.1 Case Study 1: Identifying Poor Households

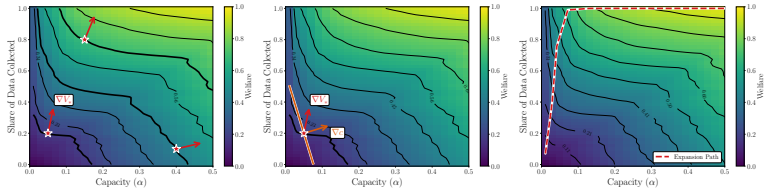
A standard objective is to maximize the number of poor households reached by the program,

$$u(y, a) = a \cdot \mathbf{1}\{y \leqslant \bar{y}\}$$

where $\bar{y}$ denotes the poverty line and $a \in \{0, 1\}$ denotes the targeting decision. Using the international poverty line of $3 per day, roughly 33% of households fell below this threshold in 2015 [World Bank, 2026a]. For now, we assume a benefit size of $100 USD PPP (purchasing power parity), corresponding to an 8% increase in yearly consumption for the average household in 2015.

(Q1) Inspecting the local rate of improvement for this design universe (Figure 3(a)), we find that a small improvement in capacity tends to far outweigh the impact of a small increase in the share of data collected. For a planner deciding where to invest at the margin, welfare is far more responsive to capacity expansion than to data collection in this setting. **(Q2)** To decide whether a program configuration is optimal, we also need to account for costs. Given the survey cost of $13 PPP per household and a transfer size of $100 PPP [Aiken et al., 2023], capacity expansion is roughly eight times more expensive per household than data collection. The natural question is whether a typical anti-poverty program — calibrated here at 20% data

[2]To simulate households without scorecard data being excluded from individual targeting, in our setting we assign these households the population-mean, corresponding to the planner having no individualized information about them. Appendix B.4 considers an alternative implementation in which households without data are treated as ineligible for the program and placed at the end of the allocation queue.



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 4: **Case Study 2.** Welfare surface for a hypothetical poverty targeting program in Ethiopia across the share of data collected and the capacity of the program. Welfare measured as gains under CRRA utility. Data collection refers to scorecard surveys at a cost of $13 USD PPP per household; capacity investment expands the share of households receiving transfers of $100 USD PPP.

coverage [Grosh et al., 2022] and 5% capacity[3] — strikes the right balance between the two. We find that it does, meaning the higher marginal welfare of capacity is exactly offset by its higher cost (Figure 3(b)), meaning that the planner is spending the budget cost-efficiently. **(Q3)** We then optimize over the design space with budgets up to $2.7 billion PPP per year[4], tracing the expansion path that characterizes the optimal budget allocation. We find that data collection and capacity expansion should go hand in hand as the budget grows.

We find that while welfare is far more responsive to capacity than to data collection, the cost structure implies that both investments matter when building a program. Importantly, a planner operating at small capacity should not focus solely on collecting data to optimally allocate a small transfer budget, but should also invest in expanding the program itself. As we show in the next case study, however, these conclusions can change under alternative specifications of the utility function.

4.1.2 Case Study 2: CRRA Utility

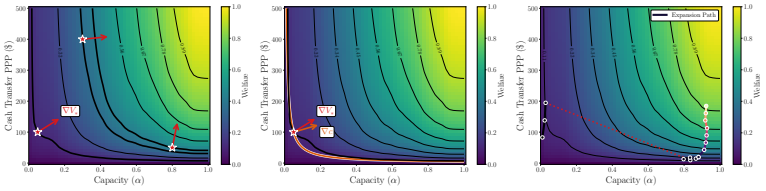
While evaluating the number of correctly identified poor households is common practice in poverty targeting, it assumes that prioritizing poorer households over less poor ones has no additional benefit, and that providing transfers to households above the poverty line has no benefit at all. The CRRA utility function [Hanna and Olken, 2018] is standard in development economics, and challenges this assumption by modeling diminishing marginal impacts of transfers as a function of household welfare,

$$u_b(w,a) = \frac{(w+ba)^{1-\rho} - w^{1-\rho}}{1-\rho}, \quad \rho > 0,\ \rho \neq 1$$

where b is the size of the cash transfer and higher ρ implies larger relative gains from targeting poorer households (see Figure 8 in Appendix). Following Hanna and Olken [2018], we use a

[3]Ethiopia's overall social protection coverage around 2015 was 21% World Bank [2026b], but the coverage of one cash transfer program is likely to be substantially below this overall coverage level.

[4]This maximum budget corresponds to 100% capacity for a $100 per year cash transfer program covering Ethiopia's 27.3 million households.



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 5: **Case Study 3.** Welfare surface for a hypothetical poverty targeting program in Ethiopia across transfer size and the capacity of the program. Welfare measured as gains under CRRA utility. Data collection is held fixed at 20%. The dotted red line in (c) indicates a discontinuous jump in the expansion path.

value of $\rho = 3$ in our empirical simulations, within the standard range 2-4 [Gandelman and Hernandez-Murillo, 2015].

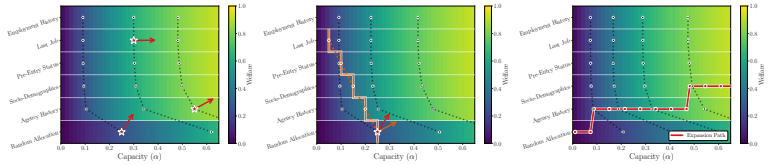
(Q1) Switching to a CRRA utility objective flips the conclusion from Case Study 1. Across larger regions of the design space, a small improvement in prediction now has a far larger impact on welfare than a small improvement in capacity (Figure 4(a)). Relative to the earlier step function, CRRA utility rewards finer distinctions among households both below and above the poverty line, making accurate targeting more valuable. **(Q2)** The same configuration as in Case Study 1 is no longer locally cost-efficient (Figure 4(b)). The planner could now improve welfare at the same total cost by rebalancing toward additional data collection and spending less on program capacity. **(Q3)** Prediction and capacity expansion no longer go hand in hand. The expansion path under the same cost structure (Figure 4(c)) shows that much of the initial budget should be spent directly on data collection.

Even though both objectives prioritize transfers to poor households, they have drastically different welfare surfaces. They flip the planner's optimal strategy from expanding capacity and data collection together to prioritizing data collection first. This points to a perhaps underappreciated lesson: choices about how to evaluate program outcomes, often made somewhat arbitrarily in practice, can imply fundamentally different design decisions and different conclusions about the value of prediction.

4.1.3 Case Study 3: Varying Transfer Sizes

So far we fixed the transfer size at $100 PPP per beneficiary. But a planner may also face a choice between writing a small number of large checks or a large number of small checks. We consider trade-offs between the intensive margin (transfer amount b) and the extensive margin (capacity α).

(Q1) Both capacity and transfer size can have large marginal impacts on welfare, and they are strongly complementary (Figure 5(a) at 20% data coverage). The marginal gain from increasing transfer size grows with program capacity, and vice versa. **(Q2)** Costs are more complex in this design space, since serving more households is more expensive when transfers are larger. This creates an interesting dynamic: in regimes where one margin has a much larger welfare impact, it also tends to be much more expensive, making the tradeoff far less obvious than it might



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 6: **Case Study 4.** Welfare surface for a hypothetical job seeker profiling system. The ordering along the feature axis is determined by greedy forward selection, cumulatively adding the group with the largest marginal welfare gain at each step. At each configuration, arrows encode the capacity increase that would be required to achieve the same welfare gain as adding the next feature group.

first appear. At the typical program configuration, for instance, the planner could improve welfare by targeting a smaller population with larger transfers (Figure 5(b)). **(Q3)** Overall, the expansion path has a discontinuity; the planner should switch between two qualitatively different strategies. At lower budgets, they should favor larger transfers to a smaller group of recipients, while at higher budgets, it becomes more valuable to provide smaller transfers to a broader share of the population.

As we saw, even a seemingly simple addition to the design space can lead to a non-trivial finding; the planner should qualitatively switch how transfers are distributed as the budget grows. But expanding the design space further to include all three levers changes the answer again, with the discontinuity disappearing and the expansion path more smoothly balancing capacity and transfer size (Figure 12 in the Appendix). Our tools enable planners to reliably make these decisions in their own domains.

4.2 Profiling of Job seekers in Germany

Public Employment Services (PES) across countries deploy predictive models to support one of their most critical tasks: the allocation of support programs to job seekers at risk of long-term unemployment (LTU) [Körtner and Bonoli, 2023, Desiere et al., 2019]. LTU incurs considerable costs for public welfare, and the support programs allocated to at-risk job seekers account for large shares of PES spending. While prediction-based profiling of job seekers is a topic of substantial policy relevance, to our knowledge the meta-design of these systems has not been systematically analyzed before. See Appendix C.2 for additional information.

Allocation Problem. Profiling tools are commonly used to predict which job seekers are at risk of long-term unemployment (12+ months in Germany, [Bach et al., 2023]) based on pre-unemployment covariates $\mathcal{X}$ such as employment and benefits history, corresponding to the utility $u(y, a) = a \cdot \mathbf{1}\{y \geqslant t_\beta\}$ from Example 3.2. Those with the highest predicted risk are prioritized for support, i.e., $\pi(x) = \mathbf{1}\{f(x) \geqslant t_\alpha\}$. To evaluate a hypothetical profiling system, we secured access to a 2% random sample of German administrative labor market records, comprising over 60 million entries of employment spells from all residents of Germany from 1970 to 2021.

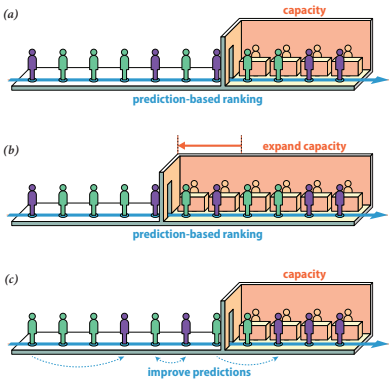


Figure 7: **(a)** Illustrates the task of identifying jobseekers at risk of long-term unemployment (purple) under a fixed capacity constraint. The employment office observes only imperfect predictions of risk, so the ranking of individuals is uncertain. **(b)** Increasing capacity expands the set of individuals who can be served. **(c)** Improving prediction sharpens the ranking, allowing limited resources to be targeted more effectively toward those truly at risk.

Design Space. From the perspective of a public employment office, we consider two design axes. The first is institutional capacity: how many job seekers can be prioritized for additional caseworker attention. The second is the quality of predictions used for targeting. In employment services, where agencies already operate on large administrative datasets, improving prediction is typically not a matter of sample size but of collecting richer information on job seekers [Desiere et al., 2019]. The dataset used in our experiments was itself constructed by linking administrative sources from the German employment system, but is not yet operationalized in agency workflows. Whether making such data available for profiling would justify the cost is a natural question; we analyze the welfare side of this tradeoff.

4.2.1 Case Study 4: Identifying Long-Term Unemployment

To operationalize the feature axis, we group available features into five categories (see Table 1). Each group contains features that broadly share a common data source and collection effort, ranging from basic socio-demographic information to longitudinal employment and agency records that require linking across administrative registries.

(Q1) We find that most of the welfare gain from prediction comes from information on past interactions with the employment agency (Figure 6(a)). Additional feature groups contribute comparatively little, suggesting that a decision rule based on agency history alone may suffice. **(Q2)** In contrast to poverty targeting, exact costs are harder to pin down, as the cost of collecting or linking features depends on institutional specifics. For illustration, we assume each feature group costs the same as expanding capacity by five percentage points (Figure 6(b)). Under these costs, investing in the first feature group is clearly worthwhile at low capacity, but once the most informative groups are in place, cost-adjusted returns to further data collection are small.

The welfare surface provides the basis for such comparisons once institutional costs are known. **(Q3)** Tracing an expansion path under this cost structure (Figure 6(c)), the optimal strategy invests early in features to avoid random allocation, then shifts toward expanding capacity.

Taken together, these results suggest that across the administrative data sources available to the agency, a (simple) decision rule based on past unemployment history captures nearly all the welfare-relevant signal. Whether qualitatively different data sources, such as structured interviews or questionnaires with job seekers [Desiere et al., 2019], could change this picture is something the same analysis could directly evaluate once such data becomes available.

Acknowledgements

UFA acknowledges the support by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education. We are thankful to Chris Hays and Sendhil Mullainathan for insightful comments and discussions. We thank Nanina Föhr for helpful feedback and support with the design of the paper's illustrations. We gratefully acknowledge the Social Sciences Computing Facility (SSCF) at UC San Diego for providing computational resources. This research utilized the Social Sciences Research and Development Environment (SSRDE) cluster.

References

E. Aiken, T. Ohlenburg, and J. Blumenstock. Moving targets: When does a poverty prediction model need to be updated? In *Proceedings of the 6th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 117–117, 2023.

E. Aiken, A. Ashraf, J. Blumenstock, R. Guiteras, and A. M. Mobarak. Scalable Targeting of Social Protection: When Do Algorithms Out-Perform Surveys and Community Knowledge? Technical report, National Bureau of Economic Research, 2025.

V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias. Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review*, 102(4):1206–1240, 2012.

S. Athey and S. Wager. Policy Learning With Observational Data. *Econometrica*, 89(1):133–161, 2021.

R. L. Bach, C. Kern, H. Mautner, and F. Kreuter. The impact of modeling decisions in statistical profiling. *Data & Policy*, 5:e32, 2023.

A. Banerjee, R. Hanna, B. A. Olken, and S. Sumarto. The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in Indonesia. *Journal of Public Economics Plus*, 1:100001, 2020.

N. Barda, D. Riesel, A. Akriv, J. Levy, U. Finkel, G. Yona, D. Greenfeld, S. Sheiba, J. Somer, E. Bachmat, et al. Developing a covid-19 mortality risk prediction model when individual-level data are not available. *Nature Communications*, 2020.

A. Barrientos. Social assistance in low and middle income countries 2000-2015. *Research Handbook on Poverty and Inequality*, 2018.

D. Bhattacharya and P. Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012.

C. Brown, M. Ravallion, and D. Van de Walle. A poor means test? Econometric targeting in Africa. *Journal of Development Economics*, 134:109–124, 2018.

S. Casacuberta and M. Hardt. Good Allocations from Bad Estimates. *arXiv preprint arXiv:2601.05597*, 2026.

B. Cockx, M. Lechner, and J. Bollens. Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics*, 80:102306, 2023. ISSN 0927-5371.

A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery.

A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

S. Desiere, K. Langenbucher, and L. Struyven. Statistical profiling in public employment services: An international comparison. OECD Social, Employment and Migration Working Papers 224, OECD Publishing, Paris, 2019. URL https://doi.org/10.1787/b5e5f16e-en.

A. N. Elmachtoub and P. Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1): 9–26, 2022.

U. Fischer-Abaigar, C. Kern, N. Barda, and F. Kreuter. Bridging the gap: Towards an expanded toolkit for ai-driven decision-making in the public sector. *Government Information Quarterly*, 41(4):101976, 2024. ISSN 0740-624X.

U. Fischer-Abaigar, C. Kern, and J. C. Perdomo. The Value of Prediction in Identifying the Worst-Off. In *International Conference on Machine Learning*, pages 17239–17261. PMLR, 2025.

N. Gandelman and R. Hernandez-Murillo. Risk Aversion at the Country Level. *Review*, 97(1): 53–66, 2015.

U. Gentilini. Cash Transfers in Pandemic Times: Evidence, Practices, and Implications from the Largest Scale up in History. *World Bank*, 2022.

D. Goller, M. Lechner, T. Pongratz, and J. Wolff. Active labor market policies for the long-term unemployed: New evidence from causal machine learning. *Labour Economics*, 94:102729, 2025. ISSN 0927-5371.

M. Grosh, J. L. Baker, et al. Proxy means tests for targeting social programs. *Living standards measurement study (LSMS) working paper*, 118:1–49, 1995.

M. Grosh, P. Leite, M. Wai-Poi, and E. Tesliuc. *Revisiting Targeting in Social Assistance: A New Look at Old Dilemmas*. World Bank Publications, 2022.

R. Hanna and B. A. Olken. Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. *Journal of Economic Perspectives*, 32(4):201–26, 2018.

L. Hu, I. R. L. Navon, O. Reingold, and C. Yang. Omnipredictors for Constrained Optimization. In *International Conference on Machine Learning*, pages 13497–13527. PMLR, 2023.

S. Jain, K. Creel, and A. Wilson. Scarce Resource Allocations That Rely on Machine Learning Should Be Randomized. In *International Conference on Machine Learning*, pages 21148–21169. PMLR, 2024.

Á. F. Junquera and C. Kern. From rules to forests: rule-based versus statistical models for jobseeker profiling. *Journal for Labour Market Research*, 59(1):1–27, 2025.

N. Kallus. Balanced Policy Evaluation and Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8909–8920, 2018.

C. Kern, R. Bach, H. Mautner, and F. Kreuter. When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes. *ACM J. Responsib. Comput.*, 1(4), Nov. 2024.

C. Kern, U. Fischer-Abaigar, J. Schweisthal, D. Frauen, R. Ghani, S. Feuerriegel, M. van der Schaar, and F. Kreuter. Algorithms for reliable decision-making need causal reasoning. *Nature Computational Science*, 5(5):356–360, 2025.

T. Kitagawa and A. Tetenov. Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616, 2018.

J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–95, May 2015.

J. Körtner and R. Bach. Inequality-Averse Outcome-Based Matching. https://osf.io/preprints/osf/yrn4d, 2023.

J. Körtner and G. Bonoli. Predictive Algorithms in the Delivery of Public Employment Services. In *Handbook of Labour Market Policy in Advanced Democracies*, pages 387–398. Edward Elgar Publishing, 2023.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116 (10):4156–4165, 2019.

L. T. Liu, I. D. Raji, A. Zhou, L. Guerdan, J. Hullman, D. Malinsky, B. Wilder, S. Zhang, H. Adam, A. Coston, B. Laufer, E. Nwankwo, M. Zanger-Tishler, E. Ben-Michael, S. Barocas, A. Feller, M. Gerchick, T. Gillis, S. Guha, D. Ho, L. Hu, K. Imai, S. Kapoor, J. Loftus, R. Nabi, A. Narayanan, B. Recht, J. C. Perdomo, M. Salganik, M. Sendak, A. Tolbert, B. Ustun, S. Venkatasubramanian, A. Wang, and A. Wilson. Bridging Prediction and Intervention Problems in Social Systems. *arXiv preprint arXiv:2507.05216*, 2026.

A. Loxha and M. Morgandi. Profiling the unemployed: a review of OECD experiences and implications for emerging economies. *Social Protection and labor discussion paper*, SP 1424, 2014.

C. F. Manski. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4): 1221–1246, 2004.

A. Mas-Colell, M. D. Whinston, J. R. Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

T. Mashiat, P. J. Fowler, and S. Das. Who Pays the RENT? implications of Spatial Inequality for Prediction-Based Allocation Policies. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1686–1697, 2025.

L. McBride and A. Nichols. Retooling Poverty Targeting Using Out-Of-Sample Validation and Machine Learning. *The World Bank Economic Review*, 32(3):531–550, 2018.

A. Narayan and N. Yoshida. Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka. *South Asia Poverty Reduction and Economic Management*, 2005.

A. Noriega-Campero, B. Garcia-Bulle, L. F. Cantu, M. A. Bakker, L. Tejerina, and A. Pentland. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 241–251. Association for Computing Machinery, 2020.

J. C. Perdomo. The Relative Value of Prediction in Algorithmic Decision Making. In *International Conference on Machine Learning*, pages 40439–40460. PMLR, 2024.

J. C. Perdomo, T. Britton, M. Hardt, and R. Abebe. Difficult Lessons on Social Prediction from Wisconsin Public Schools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2682–2704. Association for Computing Machinery, 2025.

P. Premand and P. Schnitzer. Efficiency, Legitimacy, and Impacts of Targeting Methods: Evidence from an experiment in Niger. *The World Bank Economic Review*, 35(4):892–920, 2021.

K. Ren, Y. Byun, and B. Wilder. Decision-Focused Evaluation of Worst-Case Distribution Shift. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 3076–3093. PMLR, 2024.

H. Rudolph and M. Müntnich. "Profiling" zur Vermeidung von Langzeitarbeitslosigkeit. *Mitteilungen aus der Arbeitsmarkt-und Berufsforschung*, 34(4):530–553, 2001.

A. Schmucker and P. vom Berge. Sample of Integrated LabourMarket Biographies Regional File (SIAB-R) 1975–2021, 2023a. FDZ-Datenreport, 07/2023 (en), Nürnberg.

A. Schmucker and P. vom Berge. Factually anonymous version of the Sample of Integrated Labour Market Biographies (SIAB-Regionalfile) – Version 7521 v1. Research Data Centre of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB), 2023b. URL 10.5164/IAB.SIAB-R7521.de.en.v1.

P. Schnitzer and Q. Stoeffler. Targeting Social Safety Nets: Evidence from Nine Programs in the Sahel. *The Journal of Development Studies*, 60(4):574–595, 2024.

V. Sharma and B. Wilder. Comparing Targeting Strategies for Maximizing Social Welfare with Limited Resources. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

A. Shirali, R. Abebe, and M. Hardt. Allocation Requires Prediction Only if Inequality is Low. In *International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A. Slivkins, K. A. Sankararaman, and D. J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4633–4656. PMLR, 2023.

P. Vayanos, A. Georghiou, and H. Yu. Robust optimization with decision-dependent information discovery. *Management Science*, 2026.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

A. Wang, S. Kapoor, S. Barocas, and A. Narayanan. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsib. Comput.*, 1(1), Mar. 2024.

World Bank. Poverty and Inequality Platform. https://pip.worldbank.org, 2026a. Version 20260324_2021_01_02_PROD. Accessed on 2026-04-29.

World Bank. The Atlas of Social Protection Indicators of Resilience and Equity (ASPIRE). https://www.worldbank.org/en/data/datatopics/aspire, 2026b. Accessed May 6, 2026.

S. Zezulka and K. Genin. From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1984–2006, 2024.

A Proof of Fact 3.3

Proof. Write the constraints in the form

$$c(\theta) - B \leqslant 0, \qquad \theta_i - 1 \leqslant 0, \qquad -\theta_i \leqslant 0.$$

The full Lagrangian is

$$\widetilde{\mathcal{L}}(\theta, \lambda, \mu^+, \mu^-) = V_*(\theta) - \lambda\Big(c(\theta) - B\Big) - \sum_{i=1}^{d} \mu_i^+(\theta_i - 1) - \sum_{i=1}^{d} \mu_i^-(-\theta_i),$$

where $\lambda \geqslant 0$ is the multiplier for the budget constraint, $\mu_i^+ \geqslant 0$ is the multiplier for the upper bound $\theta_i \leqslant 1$, and $\mu_i^- \geqslant 0$ is the multiplier for the lower bound $\theta_i \geqslant 0$. By the KKT conditions, the multipliers satisfy complementary slackness:

$$\begin{aligned} \lambda\Big(c(\theta^*) - B\Big) &= 0, \\ \mu_i^+(\theta_i^* - 1) &= 0 && \text{for each } i, \\ \mu_i^- \theta_i^* &= 0 && \text{for each } i. \end{aligned}$$

They also satisfy stationarity:

$$\frac{\partial V_*}{\partial \theta_i}(\theta^*) - \lambda \frac{\partial c}{\partial \theta_i}(\theta^*) - \mu_i^+ + \mu_i^- = 0 \qquad \text{for each } i.$$

Now fix a coordinate i. If $0 < \theta_i^* < 1$, then complementary slackness implies $\mu_i^+ = \mu_i^- = 0$, and stationarity gives

$$\frac{\partial V_*}{\partial \theta_i}(\theta^*) = \lambda \frac{\partial c}{\partial \theta_i}(\theta^*).$$

If $\theta_i^* = 1$, then complementary slackness implies $\mu_i^- = 0$, and stationarity gives

$$\frac{\partial V_*}{\partial \theta_i}(\theta^*) - \lambda \frac{\partial c}{\partial \theta_i}(\theta^*) = \mu_i^+ \geqslant 0.$$

Hence

$$\frac{\partial V_*}{\partial \theta_i}(\theta^*) \geqslant \lambda \frac{\partial c}{\partial \theta_i}(\theta^*).$$

If $\theta_i^* = 0$, then complementary slackness implies $\mu_i^+ = 0$, and stationarity gives

$$\frac{\partial V_*}{\partial \theta_i}(\theta^*) - \lambda \frac{\partial c}{\partial \theta_i}(\theta^*) = -\mu_i^- \leqslant 0.$$

Hence

$$\frac{\partial V_*}{\partial \theta_i}(\theta^*) \leqslant \lambda \frac{\partial c}{\partial \theta_i}(\theta^*).$$

Combining the three cases proves the result. ∎

B Poverty Targeting

B.1 Background

Proxy means tests are one of the most popular approaches for targeting cash transfer programs and other social protection programs in low-income contexts where comprehensive administrative data on incomes or consumption expenditures are unavailable. They are currently used to target social protection programs in at least fifty low- and middle-income countries [Barrientos, 2018]. The accuracy of proxy means tests in practice for identifying poor households have been studied in many studies across Africa Brown et al. [2018], Schnitzer and Stoeffler [2024], McBride and Nichols [2018], Asia Narayan and Yoshida [2005], and Latin America Noriega-Campero et al. [2020]. Other work has assessed the fairness of PMTs relative to other targeting approaches Noriega-Campero et al. [2020], susceptibility to intentional data misreporting and collusion Banerjee et al. [2020], and acceptability to data subjects Premand and Schnitzer [2021], Alatas et al. [2012], but little previous work has addressed the benefit of investing in PMT coverage and accuracy in comparison relative to investment in allocating resources to the cash transfer programs and other social programs targeted with PMTs. Aiken et al. [2025] provide an initial framework for such cost-benefit trade-offs using data on targeting cash transfers in Bangladesh, we expand this framework in this paper to account for additional policy levers and data availability constraints.

B.2 Data

The data for our simulations of cash transfer program are from Ethiopia's 2015 Living Standards Measurement Survey. The survey data covers 4,954 nationally representative households. We use data from 4,694 households that have complete information on consumption expenditures, asset possession, housing quality, and household demographics. We convert consumption expenditures data to yearly total consumption expenditures measured in USD PPP using the PPP exchange rate for 2015 from the World Bank.

B.3 Predictive Modeling

The target of our predictive model is log-transformed total household consumption expenditures. The features are 54 variables commonly included in proxy means tests in low-income countries, including:

- Asset possession, such as as whether the household owns radio, TV, satellite dish, sofa, bicycle, car, or bed
- Housing quality, such as how many rooms the dwelling has, the construction material of the roof and floor, and the primary materials used for lighting and cooking
- Household demographics, such as the number of members, number of children, gender and age of the household head, and education and literacy levels of the household head
- The region of residence of the household

All continuous feature variables are winsorized at the 99th percentile and normalized to a 0-1 range. All categorical variables are one hot encoded.

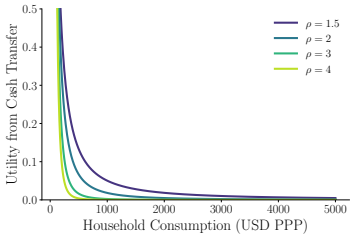
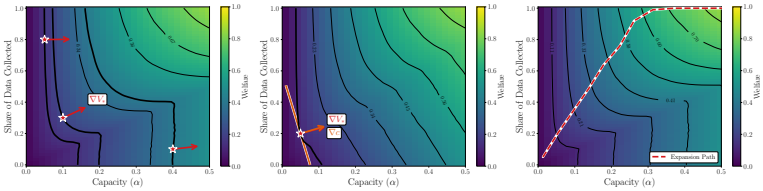


Figure 8: Normalized utility gain from a fixed $100 PPP transfer under CRRA utility for different values of ρ. Larger values of ρ place relatively more weight on transfers to lower-consumption households.



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 9: Same as Figure 3, but households without scorecard data are placed at the end of the allocation queue rather than assigned the population mean. See Section B.4 for details.

We train the predictive model on a random 75% of households and produce predictions on the remaining 25%. We repeat all our experiments for poverty targeting over 100 random train-test splits and report the average result. Our predictive model is a LASSO regression with regularization strength determined via three-fold cross validation.

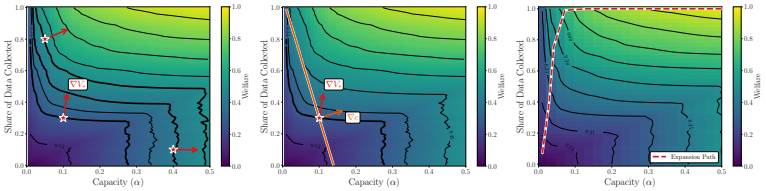
B.4 Simulating test-time data availability

Several of our experiments involve simulating lower test-time data availability – scenarios where for some households in the test set feature data is not available. We simulate this by removing predictions for these households from the test set; households without predictions receive the population mean as estimate and are targeted at random.

We also present results for an alternative implementation in which households without scorecard data are treated as ineligible for the program and placed at the end of the allocation queue.

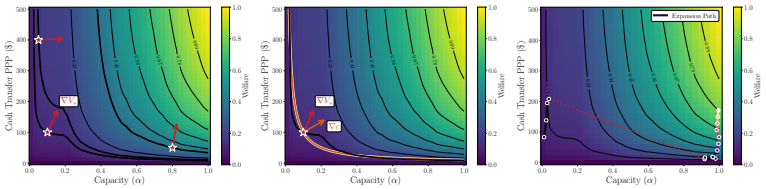
B.5 Replication

All code for the poverty targeting simulations is available, including cleaning the survey data to prepare it for predictive modeling. The survey data can be downloaded from the LSMS program at `https://microdata.worldbank.org/index.php/catalog/2783`.



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 10: Same as Figure 4, but households without scorecard data are placed at the end of the allocation queue rather than assigned the population mean. See Section B.4 for details.



(a) Local Improvements (Q1) (b) Local Optimality (Q2) (c) Expansion Path (Q3)

Figure 11: Same as Figure 5, but households without scorecard data are placed at the end of the allocation queue rather than assigned the population mean. See Section B.4 for details.

B.6 Compute Resources

All experiments (poverty-targeting and employment office case studies) were conducted on a consumer laptop. No GPU resources were used. Individual experimental runs (corresponding to individual figures in the paper) completed in under 10 minutes each, with most finishing in seconds. The total compute required to reproduce all reported results is under one hour on comparable hardware.

C Identifying Long-Term Unemployment in Germany

C.1 Background

Algorithmic profiling of job seekers in the delivery of support measures are debated and implemented by Public Employment Services (PES) in various countries [Loxha and Morgandi, 2014, Körtner and Bonoli, 2023]. In Germany, the introduction of algorithmic decision-support tools has been discussed since the early 2000s [Rudolph and Müntnich, 2001]. A common narrative in these discussions is the hope to improve efficiency over rule- or case-worker-based profiling workflows, supported by comparative studies that report higher accuracy of algorithmic profiling relative to human case-workers in predicting long-term unemployment risks [Desiere et al., 2019, Junquera and Kern, 2025]. Yet, German PES currently rely on case-

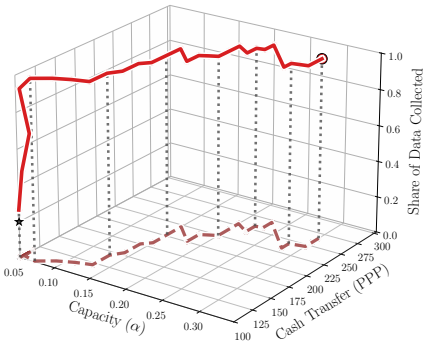


Figure 12: Expansion path for optimizing over program capacity, transfer size, and data collection starting at a transfer size of $100 PPP, 5% program capacity and 20% of poverty score cards collected. The red curve traces the welfare-maximizing design as the available budget increases.

worker-based profiling, motivating our case study as an illustrative investigation of how different profiling regimes could support employment services in Germany.

Zezulka and Genin [2024] study the link between LTU risk predictions and targeting decisions with data from the Swiss PES. In the targeting context, Cockx et al. [2023] and Goller et al. [2025] exploit effect heterogeneity of active labour market programs (ALMPs) to find optimal allocations of programs to job seekers. Körtner and Bach [2023] demonstrate how inequality-averse allocation of ALMPs to job seekers could mitigate historical differences in group-specific unemployment risks. Yet, a common thread in these works is that most parameters of the allocation problem are treated as given, i.e, they focus on modeling solutions rather than asking how outcomes could be improved with an expanded set of actions, such as additional data collection or increased screening capacity.

C.2 Data

We make use of a comprehensive dataset on German jobseekers. The Sample of Integrated Labour Market Biographies (SIAB) is provided via a Scientific Use File from the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) [Schmucker and vom Berge, 2023b]. The SIAB is a 2% sample of all governmental labor market records in Germany, comprising approximately 60 million entries spanning 1970 to 2021. It collects information on employment status, jobseeker demographics, employment and benefits histories, and participation in active labor market programs, from a range of

different data sources across the German employment services (e.g., the Employee History (BeH), Benefit Recipient History (LeH) and Unemployment Benefit II Recipient History (LHG) are separate data products within the BA). Further details on its content, sampling procedure, and anonymization can be found in the accompanying data report [Schmucker and vom Berge, 2023a].

C.3 Predictive Modeling

We construct a prediction task for long-term unemployment, predicting the number of months a job seeker will remain unemployed at entry into unemployment. Following Bach et al. [2023], Kern et al. [2024], Fischer-Abaigar et al. [2025], we use the same spell aggregation procedure to construct covariates capturing sociodemographic information, labor market and benefits history, and characteristics of the last job held. This leads to 56 numerical and 24 categorical variables; the latter are one-hot encoded. To simulate deployment conditions in real employment offices, we use records from 2010 and 2011 for training, data from 2012 for validation, and data from 2015 for testing. We train a gradient boosting model (CatBoost) on this temporal split. All policy lever comparisons in this paper are conducted on the test set ($n = 86,692$).

C.4 Replication

We provide full replication code for the design space analysis. Due to data sensitivity, the underlying records are not publicly available but can be applied for at the Research Data Centre (FDZ) of the Institute for Employment Research (IAB).

D Societal Impact

This work addresses the design of prediction-based allocation systems in public institutions, where algorithmic choices directly affect access to social goods. Our framework helps practitioners understand when prediction improvements are worthwhile relative to other investments, enabling context-specific comparisons that can inform resource allocation decisions. The framework adopts the perspective of a social planner optimizing a specified objective. This carries limitations. First, the analysis requires specifying utility functions and cost structures, excluding considerations that resist quantification, such as procedural fairness. Second, institutional objectives may diverge from the welfare of affected populations; our toolkit can inform institutional decision-making but does not resolve this tension.

Table 1: Feature groups

Group	Feature
Socio-demographics & residence	Age
	Gender
	Educational background
	State of residence
	German nationality
	Commuting status and residential moves
Pre-entry status	Employment in six weeks before unemployment entry
	Benefit receipt in six weeks before entry
	Registered job-seeking in six weeks before entry
	Registered with public employment office for other reasons in six weeks before entry
	Subsidized employment in six weeks before entry
	Missing information for the six weeks before entry
Last job	Availability of previous-job information
	Duration of last job
	Parallel employment during last job
	Type of last job
	Part-time status of last job
	Daily wage in last job
	Occupational skill level of last job
	Fixed-term and temporary-agency status of last job
	Industry of last job
Employment history	Number of previous jobs
	Number of previous establishments
	Total and mean duration of previous employment
	Duration in marginal, full-time, parallel, fixed-term, and temporary-agency employment
	Duration worked across industries
	Time since first employment
	Time since last contact with labor market
Agency history	Previous unemployment and job-search history
	Benefit receipt history
	Participation in active labor market programs
	Subsidized employment history
	Time since last job search

Part III

Outlook

8 Outlook

Across this thesis, we have argued that in order to make productive use of predictive algorithms in the public sector, we cannot evaluate a model in isolation, but only as one component of the larger institutional context in which it operates. We have developed the argument from two complementary angles. The first asks whether the assumptions underlying a system's technical implementation actually hold in its deployment setting; if they do not, high predictive accuracy alone is of little use, since, the system may be estimating something misaligned with the policy goal being pursued, or relying on a data distribution that will change after deployment. The second asks whether improving prediction is worth it at all: once a predictor is understood as one component of a resource-constrained system, its value is contextual, and should be weighed against the other investments a planner could make instead.

The second part of the thesis operates mostly in well-specified settings, effectively assuming away the complications of the first. We assumed, for example, that the planner can supply a utility function and a cost structure, and that the prediction problem is reasonably well-posed. This shows that the value of prediction is contextual even in a setting generally favorable to it, reiterating that prediction is a means to an end rather than an end in itself. It is also a limitation, and points to future work. The tools we developed could be extended to make exploratory analysis easier, letting a planner sweep across assumptions and parameter ranges and visualize how the welfare surface changes (in the spirit of multiverse analysis (Simson, Pfisterer, et al. 2024)), and account, where possible, for uncertainty over the inputs, since costs and other primitives are often not known exactly.

The same analysis could be broadened to other objectives, such as max-min welfare, where investments may matter more for some subgroups than others; to other allocation problems, such as outcome-based matching; and to richer cost structures that distinguish fixed from variable costs and amortize one-time investments. Many such problems also carry temporal and spatial structure. For example, the value of prediction may shift across employment offices, or model accuracy may decrease over time, as in poverty targeting, where surveys have to be re-run and one must decide how often to collect new data.

In theorizing the evaluation of prediction in a public institutional context, the thesis points to a broader agenda. The literature on algorithmic systems in the public sector has grown rapidly, yet the field still lacks a shared evaluation framework adequate to the complexity of the settings these systems actually operate in, though a concerted effort is currently underway (Liu et al. 2026). Developing such a toolkit is, of necessity, an interdisciplinary effort that also requires institutional knowledge to specify what a system is

actually supposed to achieve and how it fits into the organization around it. For example, further work might address the fact that public-sector institutions are rarely the simple end-to-end pipelines that much of the evaluation literature implicitly assumes: in practice, a prediction feeds into a sequence of interdependent decisions – routing choices, caseworker interactions, referral pathways – each with its own constraints and actors. Evaluation frameworks appropriate for these settings will need to reckon with such larger pipelines and their contingencies.

The value of prediction becomes harder still to evaluate when it is mediated by a caseworker or administrator. The evidence on how systems work with humans is mixed. In some cases the combination performs worse than either the human or the model alone (Vaccaro et al. 2024). In others it helps, with the human acting as a useful check on the model (De-Arteaga et al. 2020). And in general, how caseworkers will actually use a tool is hard to anticipate; it depends on their background (Cheng and Chouldechova 2022) and the incentives they face, which may reinforce tendencies such as "cream-skimming", the selection of easier cases (Lipsky 2010). Ultimately, the aim should be to build systems that support better caseworker decisions in more ways than simply handing them a risk score.

In recent years, large language models (LLMs) have emerged as a qualitatively different kind of system – one that may also offer new possibilities for supporting human decision-makers. They are already being deployed widely across public administration. Their appeal to public institutions is intuitive: as systems that operate naturally on text, they fit readily into administrative processes built around forms, reports, and written regulations. We should be cautious about deploying them as risk predictors, as they often produce poorly calibrated scores (Cruz et al. 2024), but they may help with other tasks. Caseworkers consistently cite paperwork, documentation, and keeping pace with new regulations as primary bottlenecks. Deploying LLMs to absorb some of that burden, by doing tasks such as retrieving relevant information or checking forms could free up capacity for the kinds of judgment that remain harder to automate.

However, LLMs also introduce new evaluation challenges, and ones that are in some respects more pressing than those posed by classical predictive systems. In a sense, LLMs lower the barrier to deployment. They can be used more readily out of the box, and require less upfront specification by an institution – there is no need, for example, to commit to a prediction target before proceeding. However, the reduced need for upfront goal specification may make institutions more willing to deploy such systems without the careful evaluation that deployment warrants. Developing adequate evaluation frameworks for LLMs in public administration will be no less challenging than for traditional systems (Anthis et al. 2025; Wallach et al. 2025). How to evaluate such systems rigorously, and in context of bureaucratic institutions, is among the more pressing questions the field now faces.

Algorithmic systems in the public sector are neither simply to be welcomed nor resisted. In any case, the genie is unlikely to go back in the bottle, and arguably it should not, given the potential gains in efficiency and scale that motivate their adoption. But the

fact that an institution can deploy an algorithmic system does not mean it should; what is required is an honest assessment of the value it actually brings. And to speak of value at all, we have to understand the institutional context the system enters, and stay attuned to the "costs of categorization" we may have to pay (Lipsky 2010, p. 198). Some are opportunity costs, incurred when prediction proves not to be the best way forward. Others are harder to see, and accordingly easy to underweight, as when a system improves what is readily measured while quietly eroding what is not, such as the quality of service. In one way, recognizing where systems are likely to go wrong is also what lets us identify settings more amenable to automation in the first place. Automating an existing fraud-detection system, for instance, creates a very different failure case than building a new system to help people claim benefits they are already entitled to, the latter a more forgiving environment in which to develop one. Quantification may be a "way of making decisions without seeming to decide" as Porter contends (Porter 1995, p. 8) – but it can also prompt us to ask more deeply what we want our institutions to achieve in the first place. The very effort to deploy a system and assess its value holistically opens space for an institution to articulate what it is actually trying to do, the goals it pursues, the tradeoffs it accepts, and the costs it is willing to bear.

Bibliography

Abdulkadiroğlu, A. and T. Sönmez (2003). "School Choice: A Mechanism Design Approach". *American Economic Review* 93:3, pp. 729–747. DOI: 10.1257/000282803322157061.

Achterhold, E., M. Mühlböck, N. Steiber, and C. Kern (2025). "Fairness in Algorithmic Profiling: The AMAS Case". *Minds and Machines* 35:1, p. 9. ISSN: 1572-8641. DOI: 10.1007/s11023-024-09706-9.

Agarwal, P. K. (2018). "Public Administration Challenges in the World of AI and Bots". *Public Administration Review* 78:6, pp. 917–921. DOI: 10.1111/puar.12979.

Aiken, E., A. Ashraf, J. Blumenstock, R. Guiteras, and A. M. Mobarak (2025). *Scalable Targeting of Social Protection: When Do Algorithms Out-Perform Surveys and Community Knowledge?* Working Paper 33919. National Bureau of Economic Research. DOI: 10.3386/w33919.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022). "Machine Learning and Phone Data Can Improve Targeting of Humanitarian Aid". *Nature* 603:7903, pp. 864–870. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04484-9.

Aiken, E., T. Ohlenburg, and J. Blumenstock (2023). "Moving Targets: When Does a Poverty Prediction Model Need to Be Updated?" In: *Proceedings of the 6th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*. Compass '23. Association for Computing Machinery, New York, NY, USA, p. 117. ISBN: 979-8-4007-0149-8. DOI: 10.1145/3588001.3609369.

Alford, J. and B. W. Head (2017). "Wicked and Less Wicked Problems: A Typology and a Contingency Framework". *Policy and Society* 36:3, pp. 397–413. DOI: 10.1080/14494035.2017.1361634.

Alkhatib, A. and M. Bernstein (2019). "Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Chi '19. Association for Computing Machinery, New York, NY, USA, pp. 1–13. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300760.

Allhutter, D., F. Cech, F. Fischer, G. Grill, and A. Mager (2020). "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective". *Frontiers in Big Data* 3. ISSN: 2624-909X. DOI: 10.3389/fdata.2020.00005.

Allhutter, D., A. Mager, F. Cech, F. Fischer, and G. Grill (2020). *Der AMS-Algorithmus: Eine Soziotechnische Analyse Des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*. Technical report 2020–02. DOI: 10.1553/ita-pb-2020-02.

Alur, R., M. Raghavan, and D. Shah (2024). "Human Expertise in Algorithmic Prediction". In: *Proceedings of the 38th International Conference on Neural Information Pro-*

cessing Systems. Nips '24. Curran Associates Inc., Red Hook, NY, USA. ISBN: 979-8-3313-1438-5.

Amanatidis, G., H. Aziz, G. Birmpas, A. Filos-Ratsikas, B. Li, H. Moulin, A. A. Voudouris, and X. Wu (2023). "Fair Division of Indivisible Goods: Recent Progress and Open Questions". *Artificial Intelligence* 322, p. 103965. ISSN: 0004-3702. DOI: 10.1016/j.artint.2023.103965.

Amarasinghe, K., K. T. Rodolfa, H. Lamba, and R. Ghani (2023). "Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions". *Data & Policy* 5, e5. DOI: 10.1017/dap.2023.2.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). "Machine Bias". URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 06/02/2026).

Anthis, J. R., K. Lum, M. Ekstrand, A. Feller, and C. Tan (2025). "The Impossibility of Fair LLMs". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, pp. 105–120. ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.5.

De-Arteaga, M., R. Fogliato, and A. 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, pp. 1–12. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376638.

Athey, S. (2017). "Beyond Prediction: Using Big Data for Policy Problems". *Science* 355:6324, pp. 483–485. ISSN: 00368075, 10959203. DOI: 10.1126/science.aal4321.

Athey, S., N. Keleher, and J. Spiess (2025). "Machine Learning Who to Nudge: Causal vs Predictive Targeting in a Field Experiment on Student Financial Aid Renewal". *Journal of Econometrics* 249, p. 105945. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2024.105945. URL: https://www.sciencedirect.com/science/article/pii/S0304407624002963.

Athey, S. and S. Wager (2021). "Policy Learning With Observational Data". *Econometrica* 89:1, pp. 133–161. DOI: 10.3982/ECTA15732.

Azizi, M. J., P. Vayanos, B. Wilder, E. Rice, and M. Tambe (2018). "Designing Fair, Efficient, and Interpretable Policies for Prioritizing Homeless Youth for Housing Resources". In: *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*. Ed. by W.-J. van Hoeve. Springer International Publishing, Cham, pp. 35–51. ISBN: 978-3-319-93031-2.

Baldeschi, R. C., S. Di Gregorio, S. Fioravanti, F. Fusco, I. Guy, D. Haimovich, S. Leonardi, F. Linder, L. Perini, M. Russo, et al. (2025). "Multicalibration Yields Better Matchings". *arXiv preprint arXiv:2511.11413*. arXiv: 2511.11413.

Barocas, S., M. Hardt, and A. Narayanan (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas, S. and A. D. Selbst (2016). "Big Data's Disparate Impact". *Calif. L. Rev.* 104, p. 671.

Bell, A., L. Bynum, N. Drushchak, T. Zakharchenko, L. Rosenblatt, and J. Stoyanovich (2023). "The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Association for Computing Machinery, New York, NY, USA, pp. 400–422. ISBN: 979-8-4007-0192-4. DOI: 10.1145/3593013.3594007.

Ben-Michael, E., K. Imai, and Z. Jiang (2024). "Policy Learning with Asymmetric Counterfactual Utilities". *Journal of the American Statistical Association* 119:548, pp. 3045–3058. DOI: 10.1080/01621459.2023.2300507.

Bennett, A. and N. Kallus (2020). "Efficient Policy Learning from Surrogate-Loss Classification Reductions". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 788–798.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, New York, NY. ISBN: 978-1-4419-3074-3 978-1-4757-4286-2. DOI: 10.1007/978-1-4757-4286-2.

Bhattacharya, D. and P. Dupas (2012). "Inferring Welfare Maximizing Treatment Assignment under Budget Constraints". *Journal of Econometrics* 167:1, pp. 168–196. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2011.11.007.

Brayne, S. (2017). "Big Data Surveillance: The Case of Policing". *American Sociological Review* 82:5, pp. 977–1008. DOI: 10.1177/0003122417725865.

Brioscú, A., A. Lauringson, A. Saint-Martin, and T. Xenogiani (2024). "A New Dawn for Public Employment Services: Service Delivery in the Age of Artificial Intelligence". *OECD Artificial Intelligence Papers*. DOI: 10.1787/5dc3eb8e-en.

Brodkin, E. Z. (2012). "Reflections on Street-Level Bureaucracy: Past, Present, and Future". *Public Administration Review* 72:6, pp. 940–949. DOI: 10.1111/j.1540-6210.2012.02657.x.

Bundesagentur für Arbeit (2024). *Deutscher Digitaltag: BA Setzt Künstliche Intelligenz (KI), Maschinelles Lernen Und Automatisierung Ein*. Presseinformation Nr. 26. URL: https://www.arbeitsagentur.de/presse/2024-26-deutscher-digitaltag-ba-setzt-kuenstliche-intelligenz-maschinelles-lernen-und-automatisierung-ein (visited on 02/06/2026).

Bundesministerium für Digitales und Staatsmodernisierung (2026). *Marktplatz Der KI-Möglichkeiten (MaKI)*. URL: https://www.kimarktplatz.bund.de/ (visited on 06/02/2026).

Bundesregierung (2024). *Antwort Der Bundesregierung Auf Die Kleine Anfrage Der Abgeordneten Anke Domscheit-Berg, Dr. Petra Sitte, Dr. André Hahn, Weiterer Abgeordneter Und Der Gruppe Die Linke: Einsatz Künstlicher Intelligenz Im Geschäftsbereich Der Bundesregierung*. URL: https://dserver.bundestag.de/btd/20/121/2012191.pdf (visited on 06/02/2026).

Caton, S., S. Malisetty, and C. Haas (2022). "Impact of Imputation Strategies on Fairness in Machine Learning". *J. Artif. Int. Res.* 74. ISSN: 1076-9757. DOI: 10.1613/jair.1.13197.

Charette, R. N. (2018). "Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold". *IEEE Spectrum* 24:3. URL: https://spectrum.ieee.org/trinity-nuclear-test (visited on 05/30/2026).

Cheng, L. and A. Chouldechova (2022). "Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening". *Proc. ACM Hum.-Comput. Interact.* 6:CSCW2. DOI: 10.1145/3555101.

Chouldechova, A. (2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". *Big Data* 5:2, pp. 153–163. DOI: 10.1089/big.2016.0047.

Chouldechova, A. and A. Roth (2020). "A Snapshot of the Frontiers of Fairness in Machine Learning". *Communications of the ACM* 63:5, pp. 82–89. ISSN: 0001-0782. DOI: 10.1145/3376898.

Cockx, B., M. Lechner, and J. Bollens (2023). "Priority to Unemployed Immigrants? A Causal Machine Learning Evaluation of Training in Belgium". *Labour Economics* 80, p. 102306. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2022.102306.

Collington, R. (2022). "Disrupting the Welfare State? Digitalisation and the Retrenchment of Public Sector Capacity". *New Political Economy* 27:2, pp. 312–328. DOI: 10.1080/13563467.2021.1952559.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). "Algorithmic Decision Making and the Cost of Fairness". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. DOI: 10.1145/3097983.3098095.

Coston, A., A. Kawakami, H. Zhu, K. Holstein, and H. Heidari (2023). "A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms". In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 690–704. DOI: 10.1109/SaTML54575.2023.00050.

Coston, A., E. H. Kennedy, and A. Chouldechova (2020). "Counterfactual Predictions under Runtime Confounding". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Nips '20. Curran Associates Inc., Red Hook, NY, USA. ISBN: 978-1-7138-2954-6.

Coston, A., A. Mishler, E. H. Kennedy, and A. 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, pp. 582–593. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372851.

Coyle, D. and A. Weller (2020). ""Explaining" Machine Learning Reveals Policy Challenges". *Science* 368:6498, pp. 1433–1434. DOI: 10.1126/science.aba9647.

Cruz, A. F., M. Hardt, and C. Mendler-Dünner (2024). "Evaluating Language Models as Risk Scores". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Nips '24. Curran Associates Inc., Vancouver, BC, Canada. ISBN: 979-8-3313-1438-5.

Das, S. (2022). "Local Justice and the Algorithmic Allocation of Scarce Societal Resources". *Proceedings of the AAAI Conference on Artificial Intelligence* 36:11, pp. 12250–12255. DOI: 10.1609/aaai.v36i11.21486.

Desiere, S. and L. Struyven (2021). "Using Artificial Intelligence to Classify Jobseekers: The Accuracy-Equity Trade-off". *Journal of Social Policy* 50:2, pp. 367–385. ISSN: 0047-2794, 1469-7823. DOI: 10.1017/S0047279420000203.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012). "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Itcs '12. Association for Computing Machinery, Cambridge, Massachusetts, pp. 214–226. ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255.

Elmachtoub, A. N. and P. Grigas (2022). "Smart "Predict, Then Optimize"". *Management Science* 68:1, pp. 9–26. DOI: 10.1287/mnsc.2020.3922. eprint: https://doi.org/10.1287/mnsc.2020.3922.

Elster, J. (1992). *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Russell Sage Foundation. ISBN: 978-0-87154-231-1.

Engstrom, D. F. and A. Haim (2023). "Regulating Government AI and the Challenge of Sociotechnical Design". *Annual Review of Law and Social Science* 19:Volume 19, 2023, pp. 277–298. ISSN: 1550-3631. DOI: 10.1146/annurev-lawsocsci-120522-091626.

Engstrom, D. F., D. E. Ho, C. M. Sharkey, and M.-F. Cuéllar (2020). "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies". *NYU School of Law, Public Law Research Paper* 20-54.

Eppel, R., U. Huemer, H. Mahringer, and L. Schmoigl (2026). "Algorithmic Profiling: Effective Tool for Targeting Active Labour Market Policies?" *LABOUR*. DOI: 10.1111/labr.70013.

Espeland, W. N. (1997). "Authority by the Numbers: Porter on Quantification, Discretion, and the Legitimation of Expertise". *Law & Social Inquiry* 22:4, pp. 1107–1133. ISSN: 08976546, 17474469.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., NY, United States. ISBN: 978-1-250-07431-7.

Eurofound (2025). *Austrian Court (BVwG) Rules Public Employment Service (AMS) Algorithm Complies with GDPR*. Dublin. URL: https://apps.eurofound.europa.eu/platformeconomydb/austrian-court-bvwg-rules-public-employment-service-ams-algorithm-complies-with-gdpr-110267 (visited on 05/25/2026).

Evans, T. and J. Harris (2004). "Street-Level Bureaucracy, Social Work and the (Exaggerated) Death of Discretion". *The British Journal of Social Work* 34:6, pp. 871–895. ISSN: 00453102, 1468263X. DOI: 10.1093/bjsw/bch106.

Fernandez-Loria, C. and F. Provost (2022a). "Causal Classification: Treatment Effect Estimation vs. Outcome Prediction". *Journal of Machine Learning Research* 23:59, pp. 1–35. URL: http://jmlr.org/papers/v23/19-480.html.

– (2022b). "Causal Decision Making and Causal Effect Estimation Are Not the Same… and Why It Matters". *INFORMS Journal on Data Science* 1:1, pp. 4–16.

Fischer-Abaigar, U., E. Aiken, C. Kern, and J. C. Perdomo (2026). "On the Meta-Design of Allocation Problems". *arXiv preprint arXiv:2602.08786*. arXiv: 2602.08786.

Fischer-Abaigar, U., C. Kern, N. Barda, and F. Kreuter (2024). "Bridging the Gap: Towards an Expanded Toolkit for AI-driven Decision-Making in the Public Sector". *Government Information Quarterly* 41:4, p. 101976. ISSN: 0740624X. DOI: 10.1016/j.giq.2024.101976.

Fischer-Abaigar, U., C. Kern, and J. C. Perdomo (2025). "The Value of Prediction in Identifying the Worst-Off". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. PMLR, pp. 17239–17261. URL: https://proceedings.mlr.press/v267/fischer-abaigar25a.html.

Foster, D. P. and R. V. Vohra (1997). "Calibrated Learning and Correlated Equilibrium". *Games and Economic Behavior* 21:589, pp. 40–55.

Frauen, D., V. Melnychuk, and S. Feuerriegel (2024). "Fair Off-Policy Learning from Observational Data". In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. JMLR.org, Vienna, Austria.

Freedman, R., J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer (2020). "Adapting a Kidney Exchange Algorithm to Align with Human Values". *Artificial Intelligence* 283, p. 103261. ISSN: 0004-3702. DOI: 10.1016/j.artint.2020.103261.

Friedman, B. and H. Nissenbaum (1996). "Bias in Computer Systems". *ACM Transactions on Information Systems (TOIS)* 14:3, pp. 330–347.

Gallagher, P. and R. Griffin (2023). "(In) Accuracy in Algorithmic Profiling of the Unemployed – An Exploratory Review of Reporting Standards". *Social Policy and Society*, pp. 1–14. ISSN: 1474-7464, 1475-3073. DOI: 10.1017/S1474746423000428.

Goller, D., M. Lechner, T. Pongratz, and J. Wolff (2025). "Active Labor Market Policies for the Long-Term Unemployed: New Evidence from Causal Machine Learning". *Labour Economics* 94, p. 102729. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2025.102729. URL: https://www.sciencedirect.com/science/article/pii/S0927537125000533.

Guerdan, L., A. Coston, K. Holstein, and Z. S. Wu (2023). "Counterfactual Prediction under Outcome Measurement Error". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Association for Computing Machinery, New York, NY, USA, pp. 1584–1598. ISBN: 979-8-4007-0192-4. DOI: 10.1145/3593013.3594101.

Guha, S., F. A. Khan, J. Stoyanovich, and S. Schelter (2024). "Automated Data Cleaning Can Hurt Fairness in Machine Learning-Based Decision Making". *IEEE Transactions on Knowledge and Data Engineering* 36:12, pp. 7368–7379. DOI: 10.1109/TKDE.2024.3365524.

Guo, E., G. Geiger, and J.-C. Braun (2025). "Inside Amsterdam's High-Stakes Experiment to Create Fair Welfare AI". *MIT Technology Review*. URL: https://www.technologyreview.com/2025/06/11/1118233/amsterdam-fair-welfare-ai-discriminatory-algorithms-failure/ (visited on 06/02/2026).

Hadwick, D. and S. Lan (2021). "Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany". *World Tax Journal* 13:4, pp. 609–645.

Hammond, R. J. (1966). "Convention and Limitation in Benefit-Cost Analysis". *Nat. Res. J.* 6, p. 195.

Hanna, R. and B. A. Olken (2018). "Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries". *Journal of Economic Perspectives* 32:4, pp. 201–26. DOI: 10.1257/jep.32.4.201.

Harcourt, B. E. (2018). "The Systems Fallacy: A Genealogy and Critique of Public Policy and Cost-Benefit Analysis". *The Journal of Legal Studies* 47:2, pp. 419–447.

Hardt, M., E. Price, and N. Srebro (2016). "Equality of Opportunity in Supervised Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Curran Associates Inc., Red Hook, NY, USA, pp. 3323–3331. ISBN: 978-1-5108-3881-9.

Hebert-Johnson, U., M. Kim, O. Reingold, and G. Rothblum (2018). "Multicalibration: Calibration for the (Computationally-identifiable) Masses". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1939–1948. URL: https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Hernán, M. A. (2023). *Causal Inference: What If*. Ed. by J. M. Robins. Taylor & Francis, Boca Raton. ISBN: 978-1-4200-7616-5.

Hill, M. J. and P. L. Hupe (2022). *Implementing Public Policy: An Introduction to the Study of Operational Governance*. Fourth edition. SAGE, London ; Thousand Oaks, California. ISBN: 978-1-5297-2486-8.

Hu, L., I. R. Livni Navon, O. Reingold, and C. Yang (2023). "Omnipredictors for Constrained Optimization". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 13497–13527. URL: https://proceedings.mlr.press/v202/hu23b.html.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jain, S., K. Creel, and A. Wilson (2024). "Scarce Resource Allocations That Rely on Machine Learning Should Be Randomized". *arXiv preprint arXiv:2404.08592*. arXiv: 2404.08592.

Johnson, N., S. Moharana, C. Harrington, N. Andalibi, H. Heidari, and M. Eslami (2024). "The Fall of an Algorithm: Characterizing the Dynamics Toward Abandonment".

In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Association for Computing Machinery, New York, NY, USA, pp. 337–358. ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658910.

Johnson, R. A. and S. 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, pp. 1671–1682. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533223.

Jones, N. (2015). *Anne Tolley Scraps "Lab Rat" Study on Children*. New Zealand Herald. URL: https://www.nzherald.co.nz/nz/anne-tolley-scraps-lab-rat-study-on-children/C7GIGYW2467HG327FKXFRJDPEM/ (visited on 06/01/2026).

Kallus, N. and A. Zhou (2018). "Confounding-Robust Policy Improvement". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Curran Associates Inc., Red Hook, NY, USA, pp. 9289–9299.

Kern, C., R. Bach, H. Mautner, and F. Kreuter (2024). "When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes". *ACM J. Responsib. Comput.* 1:4. DOI: 10.1145/3689485.

Kern, C., U. Fischer-Abaigar, J. Schweisthal, D. Frauen, R. Ghani, S. Feuerriegel, M. van der Schaar, and F. Kreuter (2025). "Algorithms for Reliable Decision-Making Need Causal Reasoning". *Nature Computational Science* 5:5, pp. 356–360. ISSN: 2662-8457. DOI: 10.1038/s43588-025-00814-9.

Kitagawa, T., S. Sakaguchi, and A. Tetenov (2021). "Constrained Classification and Policy Learning". *arXiv preprint arXiv:2106.12886*. arXiv: 2106.12886.

Kitagawa, T. and A. Tetenov (2018). "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice". *Econometrica* 86:2, pp. 591–616. ISSN: 1468-0262. DOI: 10.3982/ECTA13288.

Kiviat, B. (2023). "The Moral Affordances of Construing People as Cases: How Algorithms and the Data They Depend on Obscure Narrative and Noncomparative Justice". *Sociological Theory* 41:3, pp. 175–200. ISSN: 07352751, 14679558. DOI: 10.1177/07352751231186797.

Kiviat, B., S. S. Greene, and H. Yoon (2026). "Exceptions in the Algorithmic Age: Evidence from the Case of Tenant Screening". *American Journal of Sociology* 131:4, pp. 868–917.

Kiyani, S., H. Hassani, G. Pappas, and A. Roth (2025). "Robust Decision Making with Partially Calibrated Forecasts". *arXiv preprint arXiv:2510.23471*. arXiv: 2510.23471.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). "Human Decisions and Machine Predictions*". *The Quarterly Journal of Economics* 133:1, pp. 237–293. ISSN: 0033-5533. DOI: 10.1093/qje/qjx032.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). "Prediction Policy Problems". *American Economic Review* 105:5, pp. 491–495. ISSN: 0002-8282. DOI: 10.1257/aer.p20151023.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores". *arXiv preprint arXiv:1609.05807*. arXiv: 1609.05807.

Koch, B. and K. Imai (2025). "Statistical Decision Theory with Counterfactual Loss". *arXiv preprint arXiv:2505.08908*. arXiv: 2505.08908.

Kotary, J., F. Fioretto, P. Van Hentenryck, and B. Wilder (2021). "End-to-End Constrained Optimization Learning: A Survey". *arXiv preprint arXiv:2103.16378*. arXiv: 2103.16378.

Kube, A. R., S. Das, and P. J. Fowler (2023). "Fair and Efficient Allocation of Scarce Resources Based on Predicted Outcomes: Implications for Homeless Service Delivery". *Journal of Artificial Intelligence Research* 76, pp. 1219–1245.

Kuppler, M., C. Kern, R. L. Bach, and F. Kreuter (2022). "From Fair Predictions to Just Decisions? Conceptualizing Algorithmic Fairness and Distributive Justice in the Context of Data-Driven Decision-Making". *Frontiers in Sociology* Volume 7 - 2022. ISSN: 2297-7775. DOI: 10.3389/fsoc.2022.883999.

Kusner, M., J. Loftus, C. Russell, and R. Silva (2017). "Counterfactual Fairness". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 4069–4079. ISBN: 978-1-5108-6096-4.

Laszkiewicz, M., I. Daunhawer, J. E. Vogt, A. Fischer, and J. Lederer (2024). "Benchmarking the Fairness of Image Upsampling Methods". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 489–517.

Laufer, B., S. Jain, A. F. Cooper, J. Kleinberg, and H. Heidari (2022). "Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Association for Computing Machinery, New York, NY, USA, pp. 401–426. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533107.

Lazar, S. (2024). "Legitimacy, Authority, and Democratic Duties of Explanation". *Oxford Studies in Political Philosophy* 10, pp. 28–56.

Lemov, R., T. Sturm, M. D. Gordin, J. L. Klein, and L. Daston (2019). *How Reason Almost Lost Its Mind: The Strange Career of Cold War Rationality*. University of Chicago Press.

Levy, K., K. E. Chasalow, and S. Riley (2021). "Algorithms and Decision-Making in the Public Sector". *Annual Review of Law and Social Science* 17:1, pp. 309–334. ISSN: 1550-3585, 1550-3631. DOI: 10.1146/annurev-lawsocsci-041221-023808.

Li, H. (2015). *Learning to Rank for Information Retrieval and Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham. ISBN: 978-3-031-01027-9 978-3-031-02155-8. DOI: 10.1007/978-3-031-02155-8.

Lipsky, M. (2010). *Street-Level Bureaucracy : Dilemmas of the Individual in Public Services*. 30th Anniversary Expanded Edition. Russell Sage Foundation, New York. ISBN: 0-87154-544-6.

Liu, L. T., I. D. Raji, A. Zhou, L. Guerdan, J. Hullman, D. Malinsky, B. Wilder, S. Zhang, H. Adam, A. Coston, B. Laufer, E. Nwankwo, M. Zanger-Tishler, E. Ben-Michael, S. Barocas, A. Feller, M. Gerchick, T. Gillis, S. Guha, D. Ho, L. Hu, K. Imai, S. Kapoor, J. Loftus, R. Nabi, A. Narayanan, B. Recht, J. C. Perdomo, M. Salganik, M. Sendak, A. Tolbert, B. Ustun, S. Venkatasubramanian, A. Wang, and A. Wilson (2026). *Bridging Prediction and Intervention Problems in Social Systems*. arXiv: 2507.05216 [cs.LG]. URL: https://arxiv.org/abs/2507.05216.

Liu, L. T., S. Wang, T. Britton, and R. Abebe (2023). "Reimagining the Machine Learning Life Cycle to Improve Educational Outcomes of Students". *Proceedings of the National Academy of Sciences* 120:9, e2204781120. DOI: 10.1073/pnas.2204781120.

Ludwig, J., S. Mullainathan, and A. Rambachan (2024). "The Unreasonable Effectiveness of Algorithms". *AEA Papers and Proceedings* 114, pp. 623–27. DOI: 10.1257/pandp.20241072.

Mandi, J., J. Kotary, S. Berden, M. Mulamba, V. Bucarey, T. Guns, and F. Fioretto (2024). "Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities". *Journal of Artificial Intelligence Research* 80, pp. 1623–1701. ISSN: 1076-9757. DOI: 10.1613/jair.1.15320.

Manski, C. F. (2004). "Statistical Treatment Rules for Heterogeneous Populations". *Econometrica : journal of the Econometric Society* 72:4, pp. 1221–1246. DOI: 10.1111/j.1468-0262.2004.00530.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2004.00530.x.

Mashiat, T., P. J. Fowler, and S. Das (2025). "Who Pays the RENT? Implications of Spatial Inequality for Prediction-Based Allocation Policies". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 8, pp. 1686–1697.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2021). "A Survey on Bias and Fairness in Machine Learning". *Acm Computing Surveys* 54:6. ISSN: 0360-0300. DOI: 10.1145/3457607.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions". *Annual Review of Statistics and Its Application* 8:Volume 8, 2021, pp. 141–163. ISSN: 2326-831X. DOI: 10.1146/annurev-statistics-042720-125902.

NYC ADS Task Force (2019). *Automated Decision Systems Task Force Report*. Technical report. City of New York. URL: https://www.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf (visited on 06/02/2026).

Obermeyer, Z., B. Powers, C. Vogeli, and Sendhil Mullainathan (2019). "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations". *Science* 366:6464, pp. 447–453. DOI: 10.1126/science.aax2342.

ORF (2025). *AMS Gewinnt Rechtsstreit Über AMS-Algorithmus*. URL: https://orf.at/stories/3404475/ (visited on 04/12/2026).

Passi, S. and S. Barocas (2019). "Problem Formulation and Fairness". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Fat* '19. Association for Computing Machinery, Atlanta, GA, USA, pp. 39–48. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287567.

Pechenizkiy, M., H. Weerts, C. de Campos, Y. Sasaki, and J. Stoyanovich (2025). "From Benchmarking to Understanding Fairml". In: *28th European Conference on Artificial Intelligence, ECAI 2025, Including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025*. IOS Press, pp. 38–45.

Perdomo, J. C. (2024). "The Relative Value of Prediction in Algorithmic Decision Making". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 40439–40460. URL: https://proceedings.mlr.press/v235/perdomo24a.html.

Perdomo, J. C., T. Britton, M. Hardt, and R. Abebe (2025). "Difficult Lessons on Social Prediction from Wisconsin Public Schools". In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '25. Association for Computing Machinery, New York, NY, USA, pp. 2682–2704. ISBN: 979-8-4007-1482-5. DOI: 10.1145/3715275.3732175.

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press. ISBN: 978-0-691-02908-5.

Pressman, J. L. and A. B. Wildavsky (1984). *Implementation: How Great Expectations in Washington Are Dashed in Oakland*. 3. ed. Oakland Project Series. Univ. of California Press, Berkeley. ISBN: 978-0-520-05331-1.

Pruss, D. (2023). "Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Association for Computing Machinery, New York, NY, USA, pp. 312–323. ISBN: 979-8-4007-0192-4. DOI: 10.1145/3593013.3593999.

Qian, M. and S. A. Murphy (2011). "Performance Guarantees for Individualized Treatment Rules". *The Annals of Statistics* 39:2, pp. 1180–1210. ISSN: 00905364, 21688966. DOI: 10.1214/10-AOS864.

Quade, E. S. (1966). *Systems Analysis Techniques for Planning-Programming-Budgeting*. Technical report P-3322. Santa Monica, CA: The RAND Corporation.

Raji, I. D., I. E. Kumar, A. Horowitz, and A. Selbst (2022). "The Fallacy of AI Functionality". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Association for Computing Machinery, New York, NY, USA, pp. 959–972. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533158.

Rambachan, A., A. Coston, and E. Kennedy (2022). "Robust Design and Evaluation of Predictive Algorithms under Unobserved Confounding". *arXiv preprint arXiv:2212.09844*. arXiv: 2212.09844.

Regulation (EU) 2024/1689 (Artificial Intelligence Act) (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L 2024/1689, pp. 1–144. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Richardson, R. (2021). "Defining and Demystifying Automated Decision Systems". *Maryland Law Review* 81, p. 785.

Rittel, H. W. and M. M. Webber (1973). "Dilemmas in a General Theory of Planning". *Policy Sciences* 4:2, pp. 155–169.

Roth, A. E. (2007). "Repugnance as a Constraint on Markets". *Journal of Economic Perspectives* 21:3, pp. 37–58.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". *Journal of Educational Psychology* 66:5, pp. 688–701.

Saleiro, P., B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani (2018). "Aequitas: A Bias and Fairness Audit Toolkit". *arXiv preprint arXiv:1811.05577*. arXiv: 1811.05577.

Saxena, D., K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha (2021). "A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare". *Proc. ACM Hum.-Comput. Interact.* 5:CSCW2. DOI: 10.1145/3476089.

Selbst, A. D., D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi (2019). "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Fat* '19. Association for Computing Machinery, New York, NY, USA, pp. 59–68. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287598.

Sharma, V. and B. Wilder (2025). "Comparing Targeting Strategies for Maximizing Social Welfare with Limited Resources". In: *International Conference on Learning Representations*. Ed. by Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu. Vol. 2025, pp. 19840–19860. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/3272bc34d94f20e8ac7246f5f956581a-Paper-Conference.pdf.

Shirali, A., R. Abebe, and M. Hardt (2024). "Allocation Requires Prediction Only If Inequality Is Low". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 45114–45153. URL: https://proceedings.mlr.press/v235/shirali24a.html.

Signé, L. (2017). "Policy Implementation–A Synthesis of the Study of Policy Implementation and the Causes of Policy Failure". *OCP Policy Center* 9, p. 26.

Simon, H. A. (1997). *Administrative Behavior*. 4th ed. Free Press, New York. ISBN: 978-0-684-83582-2.

Simson, J., A. Fabris, and C. Kern (2024). "Lazy Data Practices Harm Fairness Research". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Trans-*

parency. FAccT '24. Association for Computing Machinery, New York, NY, USA, pp. 642–659. ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658931.

Simson, J., F. Pfisterer, and C. Kern (2024). "One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1305–1320.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". *Statistical Science* 5:4, pp. 465–472. ISSN: 08834237, 21688745.

Stone, P. (2007). "Why Lotteries Are Just". *Journal of Political Philosophy* 15:3, pp. 276–295. DOI: 10.1111/j.1467-9760.2006.00274.x.

Stoyanovich, J., S. Abiteboul, B. Howe, H. V. Jagadish, and S. Schelter (2022). "Responsible Data Management". *Communications of The Acm* 65:6, pp. 64–74. ISSN: 0001-0782. DOI: 10.1145/3488717.

Sun, T. Q. and R. Medaglia (2019). "Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare". *Government Information Quarterly* 36:2, pp. 368–383. ISSN: 0740-624X. DOI: 10.1016/j.giq.2018.09.008.

Tobin, J. (1970). "On Limiting the Domain of Inequality". *The Journal of Law & Economics* 13:2, pp. 263–277. ISSN: 00222186.

Vaccaro, M., A. Almaatouq, and T. Malone (2024). "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis". *Nature Human Behaviour* 8:12, pp. 2293–2303. ISSN: 2397-3374. DOI: 10.1038/s41562-024-02024-1.

van Noordt, C. and G. Misuraca (2022). "Artificial Intelligence for the Public Sector: Results of Landscaping the Use of AI in Government across the European Union". *Government Information Quarterly* 39:3, p. 101714. ISSN: 0740-624X. DOI: 10.1016/j.giq.2022.101714.

Vieth-Ditlmann, K., A. Lorenz, and J. Botta (2025). *Ein verbindliches KI-Transparenzregister für Deutschland*. URL: https://algorithmwatch.org/de/wp-content/uploads/2025/08/250805-Policy-Papier_KI-Transparenzregister.pdf (visited on 06/02/2026).

Vredenburgh, K. (2025). "AI and Bureaucratic Discretion". *INQUIRY* 68:4, pp. 1091–1120.

Wager, S. (2025). "Causal Inference: A Statistical Learning Approach". URL: https://web.stanford.edu/~swager/causal_inf_book.pdf (visited on 05/25/2026).

Waldman, A. E. (2019). "Power, Process, and Automated Decision-Making". *Fordham L. Rev.* 88, p. 613.

Wallach, H., M. Desai, A. F. Cooper, A. Wang, C. Atalla, S. Barocas, S. L. Blodgett, A. Chouldechova, E. Corvi, P. A. Dow, J. Garcia-Gathright, A. Olteanu, N. J. Pangakis, S. Reed, E. Sheng, D. Vann, J. W. Vaughan, M. Vogel, H. Washington, and A. Z. Jacobs (2025). "Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge". In: *Proceedings of the 42nd International Conference on Machine Learn-*

ing. Vol. 267. Proceedings of Machine Learning Research. PMLR, pp. 82232–82251. URL: https://proceedings.mlr.press/v267/wallach25a.html.

Wang, A., S. Kapoor, S. Barocas, and A. Narayanan (2023). "Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms That Optimize Predictive Accuracy". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Association for Computing Machinery, New York, NY, USA, p. 626. ISBN: 979-8-4007-0192-4. DOI: 10.1145/3593013.3594030.

Wilder, B., B. Dilkina, and M. Tambe (2019). "Melding the Data-Decisions Pipeline: Decision-focused Learning for Combinatorial Optimization". *Proceedings of the AAAI Conference on Artificial Intelligence* 33:01, pp. 1658–1665. DOI: 10.1609/aaai.v33i01.33011658.

Wirtz, B. W., J. C. Weyerer, and C. Geyer (2019). "Artificial Intelligence and the Public Sector—Applications and Challenges". *International Journal of Public Administration* 42:7, pp. 596–615. DOI: 10.1080/01900692.2018.1498103. eprint: https://doi.org/10.1080/01900692.2018.1498103.

Yang, K. and J. Stoyanovich (2017). "Measuring Fairness in Ranked Outputs". In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. SSDBM '17. Association for Computing Machinery, New York, NY, USA. ISBN: 978-1-4503-5282-6. DOI: 10.1145/3085504.3085526.

Zadrozny, B. (2003). *Policy Mining: Learning Decision Policies from Fixed Sets of Data*. PhD Thesis, University of California, San Diego.

Zehlike, M., K. Yang, and J. Stoyanovich (2022a). "Fairness in Ranking, Part I: Score-Based Ranking". *ACM Computing Surveys* 55:6. ISSN: 0360-0300. DOI: 10.1145/3533379.

– (2022b). "Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems". *ACM Computing Surveys* 55:6. ISSN: 0360-0300. DOI: 10.1145/3533380.

Zhao, S., M. P. Kim, R. Sahoo, T. Ma, and S. Ermon (2021). "Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Nips '21. Curran Associates Inc., Red Hook, NY, USA. ISBN: 978-1-7138-4539-3.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". *Journal of the American Statistical Association* 107:449, pp. 1106–1118. ISSN: 0162-1459 1537-274X. DOI: 10.1080/01621459.2012.695674.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Im Rahmen dieser Arbeit wurden KI-basierte Werkzeuge, Claude (Anthropic) und ChatGPT (OpenAI), unterstützend bei der Erstellung von Programmcode sowie zur sprachlichen Überarbeitung von Textpassagen eingesetzt. Die inhaltliche Ausarbeitung und die wissenschaftliche Verantwortung liegen vollständig beim Autor.

München, den 08. Juni 2026 Unai Fischer Abaigar