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

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 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

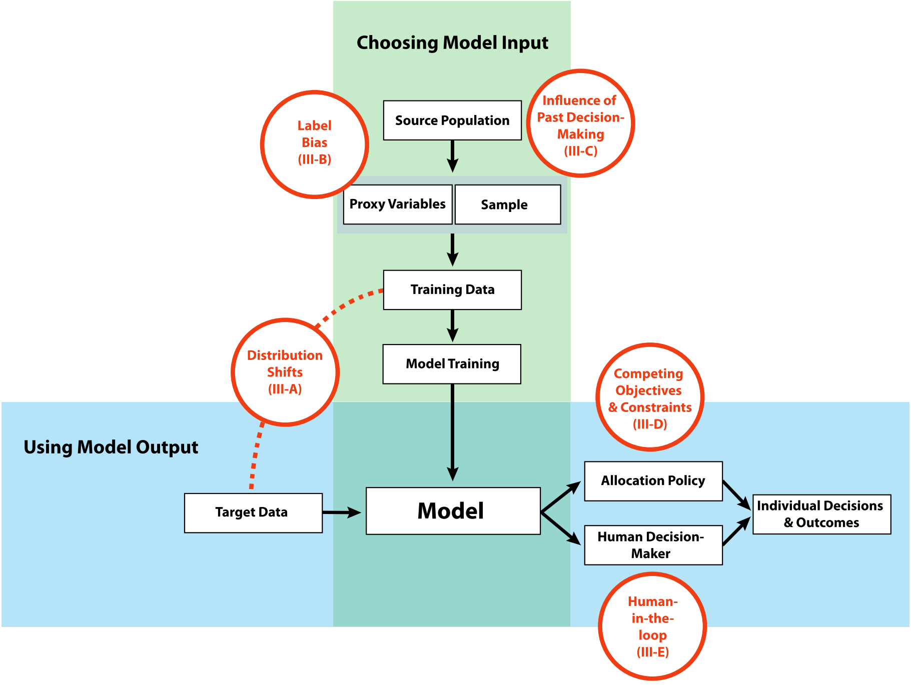


Fig. 1. Overview of the Primary Technical Challenges at the Intersection of Public Sector Decision-Making and Machine Learning. The challenges (highlighted in red) are positioned along the ML pipeline, with emphasis on data collection and model training (green) and model deployment to support decision-making (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 $\widetilde{Y}$ (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;

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

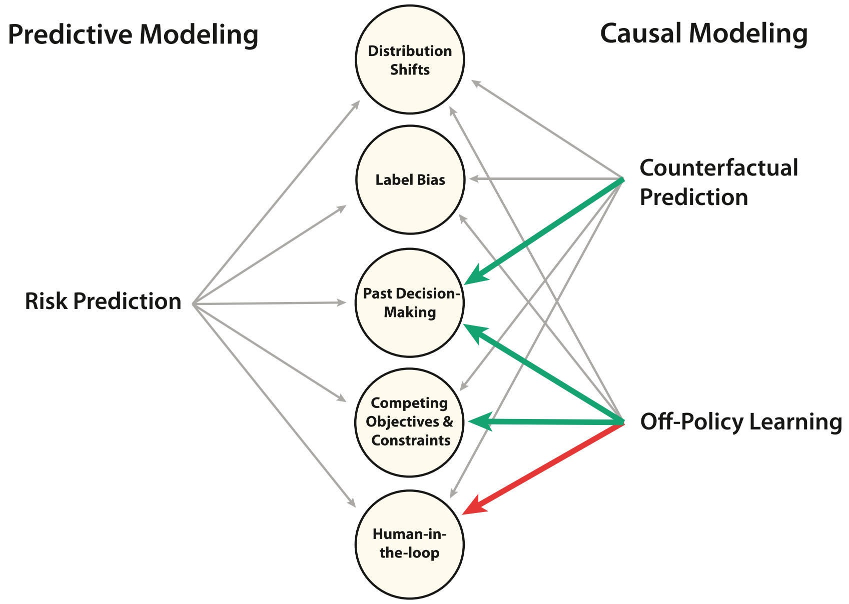


Fig. 2. Comparing ML Frameworks for Public Sector Decision-Making. Green lines indicate where a causal ML framework is potentially better at addressing a challenge, red lines highlight additional difficulties, and gray lines represent the baseline difficulty of using standard predictive modeling. (For 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 $f : \mathscr{X} \rightarrow \mathscr{Y}$ 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 $\mathscr{T} = \{0, 1\}$, 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 $\mathscr{D} = \{(X_i, T_i, Y_i)\}_{i=1}^n$ is challenging, as it is usually only possible to observe one outcome $Y_i = (1 - T_i)Y_i(0) + T_iY_i(1)$ 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 $x \in \mathscr{X}$, 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 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 $\pi : \mathscr{X} \rightarrow \{0, 1\}$, 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. $\widehat{\pi} = argmax_{\pi \in \Pi} \ \widehat{V}(\pi)$. 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 linking estimand to policy objective	→ ←	→ ←	Most assumptions for estimation Least assumptions for linking estimand to policy objective

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

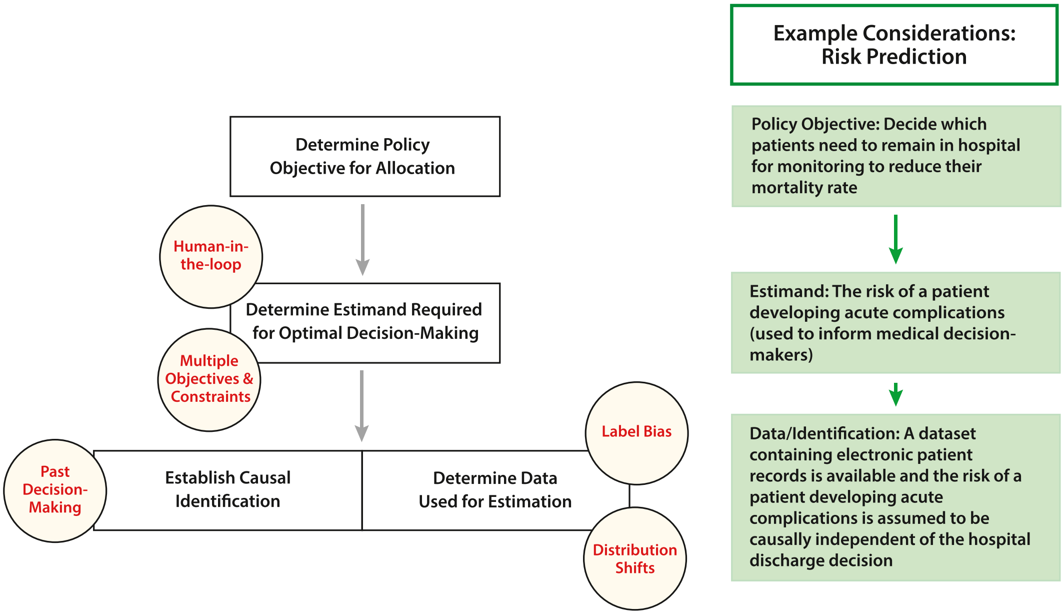
Table A1
Overview of Methodological Approaches to Address Key Challenges of ADM Systems in the Public Sector in Risk Prediction, Counterfactual Prediction, and Off-Policy Learning.

	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)

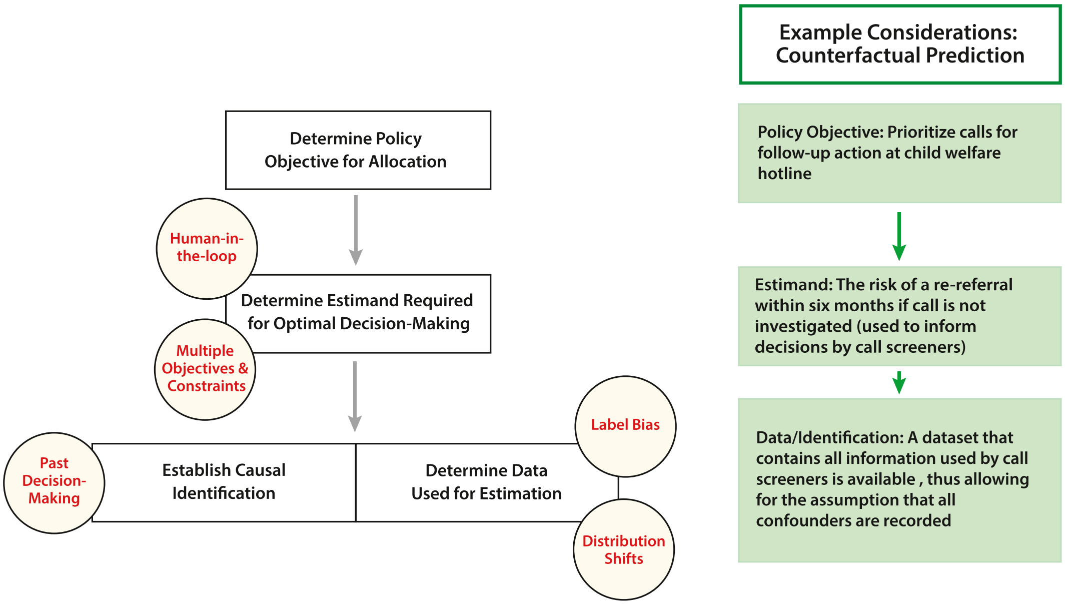

Example Considerations: Risk Prediction
Determine Policy Objective for Allocation
Human-in-the-loop
Determine Estimand Required for Optimal Decision-Making
Multiple Objectives & Constraints
Past Decision-Making
Establish Causal Identification
Determine Data Used for Estimation
Label Bias
Distribution Shifts

Policy Objective: Decide which patients need to remain in hospital for monitoring to reduce their mortality rate

Estimand: The risk of a patient developing acute complications (used to inform medical decision-makers)

Data/Identification: A dataset containing electronic patient records is available and the risk of a patient developing acute complications is assumed to be causally independent of the hospital discharge decision

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).


Example Considerations: Counterfactual Prediction
Determine Policy Objective for Allocation
Human-in-the-loop
Determine Estimand Required for Optimal Decision-Making
Multiple Objectives & Constraints
Past Decision-Making
Establish Causal Identification
Determine Data Used for Estimation
Label Bias
Distribution Shifts

Policy Objective: Prioritize calls for follow-up action at child welfare hotline

Estimand: The risk of a re-referral within six months if call is not investigated (used to inform decisions by call screeners)

Data/Identification: A dataset that contains all information used by call screeners is available , thus allowing for the assumption that all confounders are recorded

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).

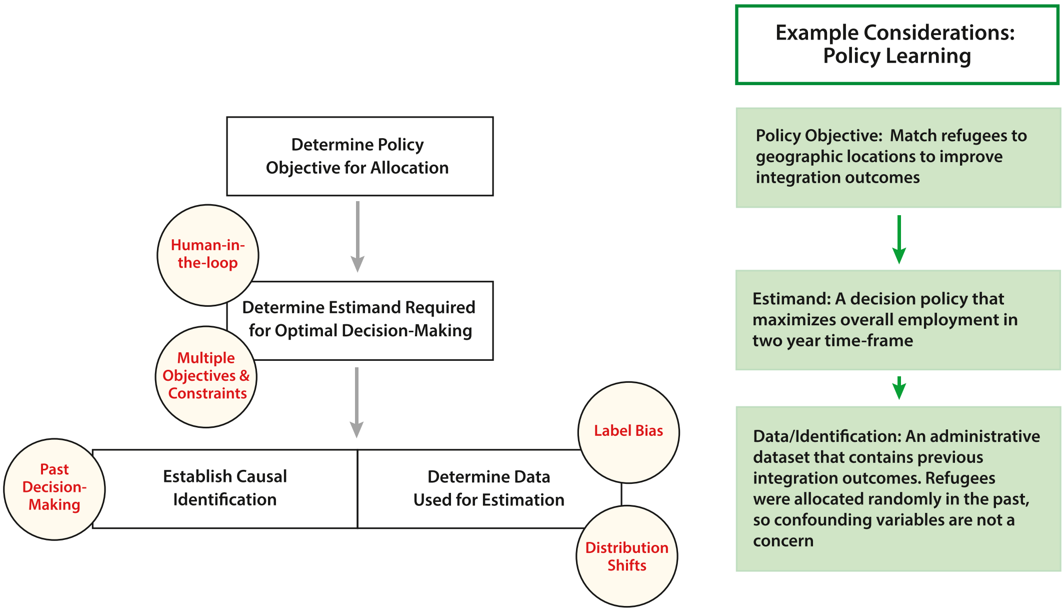

Example Considerations: Policy Learning
Policy Objective: Match refugees to geographic locations to improve integration outcomes
Estimand: A decision policy that maximizes overall employment in two year time-frame
Data/Identification: An administrative dataset that contains previous integration outcomes. Refugees were allocated randomly in the past, so confounding variables are not a concern
Determine Policy Objective for Allocation
Human-in-the-loop
Determine Estimand Required for Optimal Decision-Making
Multiple Objectives & Constraints
Past Decision-Making
Establish Causal Identification
Determine Data Used for Estimation
Label Bias
Distribution Shifts

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.

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*

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.