Bridging the Gap: Towards an Expanded Toolkit for ML-Supported Decision-Making in the Public Sector

Unai Fischer-Abaigar[1,2,*] Christoph Kern[1,2,3] Noam Barda[4]
Frauke Kreuter[1,2,3]

[1]Dept. of Statistics, LMU Munich, Germany
[2] Munich Center for Machine Learning, Germany
[3] Joint Program in Survey Methodology, University of Maryland, USA
[4] Dept. of Software and Information Systems Engineering and Dept. of Epidemiology, Biostatistics and Community Health Sciences, Ben Gurion University of the Negev, Israel

Abstract

Machine Learning (ML) 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, they still face the challenge of aligning nuanced policy objectives with the precise formalization requirements necessitated by ML models. In this paper, we aim to bridge the gap between ML model requirements and public sector decision-making by presenting a comprehensive overview of key technical challenges where disjunctions between policy goals and ML models commonly arise. We concentrate on pivotal points of the ML pipeline that connect the model to its operational environment, discussing the significance of representative training data and highlighting the importance of a model setup that facilitates effective decision-making. Additionally, we link these challenges with emerging methodological advancements, encompassing causal ML, domain adaptation, uncertainty quantification, and multi-objective optimization, illustrating the path forward for harmonizing ML and public sector objectives.

Keywords— automated decision-making, reliable artificial intelligence, public policy, causal machine learning

1 Introduction

Automated decision-making (ADM) systems are increasingly being adopted across the public sector (Algorithm Watch, 2020; Mitchell et al., 2021; Levy et al., 2021), often relying on machine learning (ML) models to address a wide array of problem domains, including critical areas such as predictive policing (Lum and Isaac, 2016), criminal justice (Angwin et al., 2016; McKay, 2020), fraud detection in government (Engstrom et al., 2020), child abuse prevention (Chouldechova et al., 2018), tax audit selection (Black et al., 2022), early warning systems in public schools (Perdomo et al., 2023), credit scoring (Kozodoi et al., 2022), profiling of jobseekers (Desiere and Struyven, 2021; Körtner and Bonoli, 2023; Bach et al., 2023), development aid (Kuzmanovic et al., 2024) and public health

*Corresponding author.

(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 et al., 2019; Coston et al., 2023; Engstrom et al., 2020; Wang et al., 2023; Levy et al., 2021). Real-world examples frequently demonstrate shortcomings, ranging from racial and gender bias to systems exhibiting poor predictive accuracy leading to flawed decision-making (Obermeyer et al., 2019; Allhutter et al., 2020; Angwin et al., 2016; Dressel and Farid, 2018; Mayer et al., 2020). Such unintended consequences 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 et al., 2023).

A growing body of literature explores the challenges and potential benefits of employing AI systems to enhance decision-making within the public sector (Sun and Medaglia, 2019; Zuiderwijk et al., 2021; Pencheva et al., 2020; Wirtz et al., 2019). 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 and Misuraca, 2022; Algorithm Watch, 2019). While these reviews cover a wide range of challenges, their focus is on institutional and social implications rather than technical challenges and methodological developments.

One central challenge in implementing algorithmic systems in the public sector is the tension between complex, ambiguous policy objectives and the explicit formalization requirements demanded by ML models (Levy et al., 2021; Mitchell et al., 2021; Amarasinghe et al., 2023; Passi and Barocas, 2019). Formulating policy objectives and decision-making invariably involves various stakeholders, frequently entangled in a political process characterized by conflicting goals (Levy et al., 2021; Coyle and Weller, 2020). Building a system that aligns with the originally intended outcome, defined by complex and ambiguous political compromises (Coyle and Weller, 2020), necessitates careful consideration of various technical choices during development, including an examination of data assumptions and the model's integration into the decision-making process. Efforts to address these challenges from a technical perspective are ongoing and focus on connecting methodological ML research with the unique demands of high-stakes decision-making. These efforts explore various subdimensions of this complex issue, including explainability (Amarasinghe et al., 2023), fairness (Mitchell et al., 2021; Körtner and Bach, 2023; Barocas et al., 2019), target variable bias (Guerdan et al., 2023b) and uncertainty (Kaiser et al., 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).

Unlike many other ML papers focused on decision-making, our paper is exclusively concerned with the issues that emerge when conducting high-stakes decision-making in the public sector. We will discuss ML decision making in a way that is adapted to this specific context by giving a detailed discussion of resource allocation problems commonly encountered in the public sector. These scenarios usually involve a set of individuals, which may refer to individual people, organizations or cases (Kuppler et al., 2022). For each individual a decision must be made regarding an appropriate intervention from a set of possible interventions. Given that many interventions involve some form of cost and that resources in the public sector are typically scarce, the decision-making process typically revolves around determining whether an individual should receive a particular good or service. For example, we may wish to distribute limited ICU beds among patients, or identify the financial reports that warrant investigation for potential fraud. ML models are then employed to either directly estimate the optimal allocation of interventions, or to predict the individual outcomes under different interventions to inform the downstream decision-making process.

This paper contributes to bridging the gap between ML and decision-making in the public sector by providing a concise yet holistic overview of central methodological decision-points. We present three modeling frameworks and illustrate their use and alignment with decision-making scenarios in the public sector: risk prediction, counterfactual prediction and policy learning. We map each modeling framework to key points at which disjunctions between policy goals and models can arise, and highlight methodological advancements from the technical literature to help avoid common

pitfalls in ADM practice. It offers a critical yet productive addition to this field by not only giving a systematic overview of central inflection points in the model-building process but also directly connecting them with promising methodological developments. This paper should thus be of use to any ML practitioner and policymaker seeking an in-depth discussion of the fundamental challenges of building models for public sector decision-making and also serves as a well-structured research guide to the interdisciplinary array of methods available to address these challenges.

We will first explore central technical challenges that occur along the ML pipeline when developing and deploying ML systems to support decision-making in the public sector (Section 2). Second, we will highlight recent methodological developments that exceed the classical supervised ML paradigm, showing promise in addressing the challenges identified (Section 3). In the discussion (Section 4), we place special emphasis on selecting an appropriate modeling approach and target estimand to effectively inform decision-making in a given context.

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 et al., 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 and 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 and 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 (Hüllermeier, 2021; Coston et al., 2023).

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 and 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 technical challenges associated with both data input and model output that are especially relevant for high-stakes decision-making in the public sector (see Figure 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).

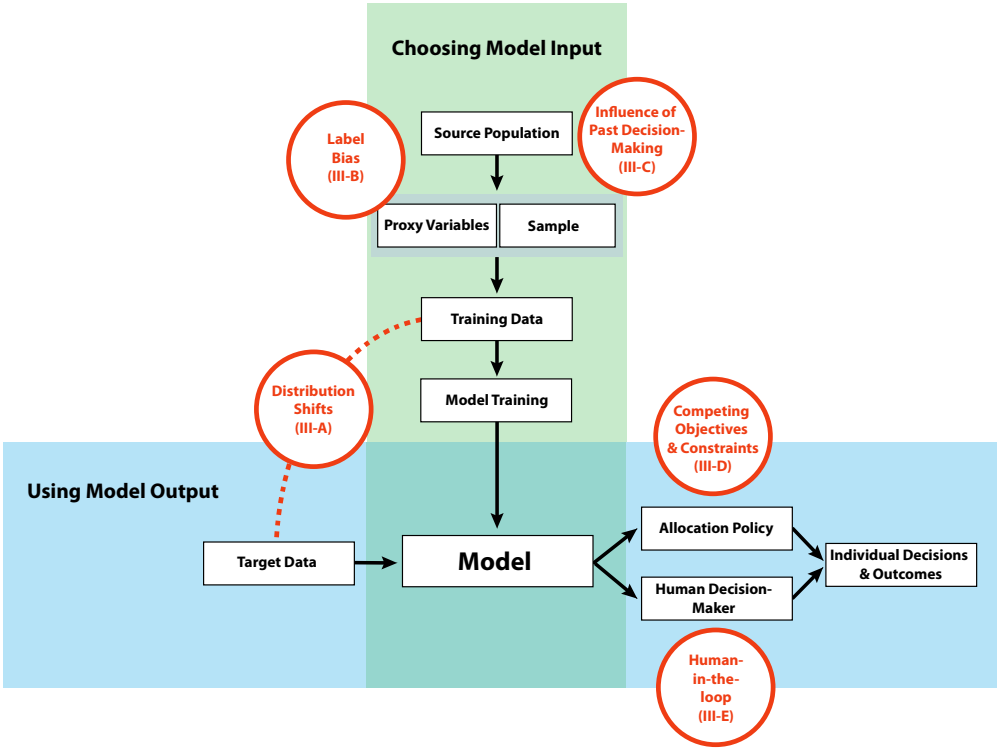

Figure 1: Overview of the Primary Technical Challenges at the Intersection of Public Sector Decision-Making and Machine Learning. The challenges (highlighted in red) are positioned along the ML pipeline, with emphasis on data collection and model training (green) and model deployment to support decision-making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

2.1 Distribution Shifts

A fundamental challenge when deploying ML models is that the data on which the system is trained and evaluated may not accurately represent the population of interest in the deployment environment (Kouw and Loog, 2018; Gruber et al., 2023). When a model is trained using source data $\mathcal{D}_S$ that is differently distributed than the target data $\mathcal{D}_T$ it will be used on in the relevant application,

the model is likely to experience a decline in performance. This phenomenon is commonly referred to as distribution or dataset shift between the joint distribution of the source data $p_{\mathcal{S}}(X, Y)$ and the target distribution $p_{\mathcal{T}}(X, Y)$ of the deployment data. Commonly investigated types of shifts follow from possible decompositions of the joint distribution $p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y)$, assuming that one component remains invariant, while the other differs (Kouw and Loog, 2018; Moreno-Torres et al., 2012). For example, shifts can occur in the covariates $p_{\mathcal{S}}(X) \neq p_{\mathcal{T}}(X)$ or in labels $p_{\mathcal{S}}(Y) \neq p_{\mathcal{T}}(Y)$, with the respective conditional distributions $p(Y|X)$ or $p(X|Y)$ remaining unchanged. Meanwhile, concept shift refers to a change in the relationship between covariates and labels, while keeping prior distributions $p(X)$ or $p(Y)$ constant.

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, 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. Similarly, deploying a model in a new problem context, for instance in a different geographic region, may also lead to a distribution shift.

Medical prediction models are instructive here, as population health and healthcare practices change dramatically over time and place, making these models particularly susceptible to distribution shifts. For example, the well-known Framingham models used for cardiovascular disease risk prediction were found to overestimate risk within the European population (Damen et al., 2019). This overestimation may be attributable to advances in available preventive treatments since the models were initially developed using data from the second half of the 20th century (Damen et al., 2019).

However, even a perfect 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). 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 et al., 2016a). 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 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 et al., 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 et al., 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 et al., 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 for the prediction target in real-world settings is rarely simple. Frequently, the true target is not directly measurable (Coston et al., 2023; Guerdan et al., 2023b; Barocas et al., 2019), as is the case with health outcomes or crime rates. This often encourages 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 et al., 2019a).

However, 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 et al., 2023b; Passi and 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 et al., 2020) and lending algorithms (Mitchell et al., 2021).

Mitigating such biases cannot be achieved by collecting more data; it demands careful consideration of the relationship between the the true label Y and the measured proxy label $\tilde{Y}$ (Gruber et al., 2023; Guerdan et al., 2023b). 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 (Vaithianathan et al., 2019; Cheng and Chouldechova, 2022).

2.3 Past Decision-Making

When developing an ADM system, we often encounter scenarios in which the available training has been influenced by past decision-making (Coston et al., 2020). For example, a model predicting the risk of jobseekers 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 et al., 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 and 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, 2020a; Kuzmanovic et al., 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 and 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 et al., 2022). However, in many scenarios it will be impossible to conduct a RCT due to resource constraints or 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 et al., 2017) or child abuse prevention (Vaithianathan et al., 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 (Rambachan et al., 2022; Lakkaraju et al., 2017), 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 et al. (2023) and Lakkaraju et al. (2017) utilize data from multiple randomly assigned 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 and 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 et al., 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 and 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 and 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 (Kozodoi et al., 2022; Black et al., 2022) or enhance model interpretability (Murdoch et al., 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 et al., 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., 2019). 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 and Weller, 2020; Levy et al., 2021).

2.5 Human-in-the-Loop

In many critical application contexts, the goal is not to entirely replace humans with ADM systems, but rather to aid human decision-makers in their tasks (Enarsson et al., 2022). Human oversight becomes particularly important in scenarios where the ML system alone can not fulfill all the necessary system requirements (Doshi-Velez and Kim, 2017), as determined by the underlying policy goals. Human quality control can be essential to ensure both system reliability and safety, especially when quantifying and anticipating model uncertainties proves challenging (Gruber et al., 2023). Additionally, as the definition of an ADM system's objective can often be difficult, human input can aid in balancing complex and hard to formalize objectives.

These issues are often reflected in legal and regulatory restrictions. For example, the current EU proposal for an AI legal framework stresses the importance of human oversight for high-risk systems in Article 14 (EU Commission, 2021). In such scenarios, humans maintain the final say in determining the right course of action, necessitating considerations of how they will interpret model outputs and incorporate them into their judgment. Generally, this requires shifting the focus from estimating the most accurate model to evaluating the consequences of providing a human decision-maker with a particular recommendation (Fernández-Loría and Provost, 2022b; Vodrahalli et al., 2022).

Studies have shown that users are often hesitant to follow recommendations of a predictive model, a phenomenon known as algorithm aversion (De-Arteaga et al., 2020; Dietvorst et al., 2018, 2015). This lies in contrast with the opposing tendency of automation bias, when humans excessively rely on a machine's suggestion (Goddard et al., 2012; De-Arteaga et al., 2020). Ongoing research into how humans interact with algorithmic decision-making systems (Chugunova and 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 (Yeomans et al., 2019; Nourani et al., 2019; Amarasinghe et al., 2023). For instance, Lebovitz et al. (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 et al., 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 and 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: From Allocation Principle to Target Estimand

Successfully constructing ML systems to inform decision-making and addressing the outlined challenges is a fundamentally difficult task that can not easily be tackled in the traditional supervised ML framework. Consequently, there have been calls to move beyond purely predictive modeling towards ML methodology that 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. However, selecting an appropriate modeling approach first requires an understanding of which target estimand can best inform decision-making for a given project.

This task is complex, as it can be challenging to clarify the intended guiding principles of an ADM system even when certain goals are already set. For example, public employment agencies often seek to allocate resources to jobseekers 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 jobseekers or the notion that high-risk jobseekers inherently deserve more support (Desiere and 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.

In practice, this discussion becomes even more complicated than the scope we can cover here, given that decision-making often extends beyond binary interventions at a specific point in time (Hüllermeier, 2021). This complexity may require alternative target estimands, especially in scenarios involving multi-valued and continuous interventions that can simultaneously target multiple risk factors or occur sequentially at different time points (Lin et al., 2021; Acharki et al., 2023; Van Geloven et al., 2020). Recent works have emphasized the importance of clearly defining the target estimand (Lundberg et al., 2021). For example, Lin et al. (2021) and Van Geloven et al. (2020) review various estimands and modeling approaches for clinical prediction models. Clear definition of the target of interest is of critical importance, context-dependent and should align with the requirements of the decision-making process.

In the following sections, we will explore classical risk prediction (Section 3.1), counterfactual prediction (Section 3.2) and policy learning (Section 3.3) as key instances of target estimands of an ADM system. We show in which context which estimand is particularly relevant, especially when the modeling of past decision-making is required (Section 2.3), and discuss different estimation strategies for each. We then present methodological advancements for each modeling approach with which the challenges described in the previous section can be addressed in a given estimation setting.

Specifically, we focus on distribution shifts to ensure robustness across deployment environments, uncertainty quantification as a key building block for generating trustworthy predictions for humans, and multi-objective optimization to manage tradeoffs between competing objectives.

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 et al., 2015; Guerdan et al., 2023b; Fernández-Loría and 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 and 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 jobseekers with the goal of allocating interventions on this basis (Allhutter et al., 2020). This model divided jobseekers 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 et al., 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 $f : \mathcal{X} \to \mathcal{Y}$ from a set of observed training data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$. 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 and Namkoong, 2021). Training models that remain valid under distribution shift often requires assumptions about the expected type of shift (David et al., 2010). As outlined in Section 2.1, we will predominantly focus on on covariate, label and concept shits (Moreno-Torres et al., 2012; Quinonero-Candela et al., 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 et al. (2022).

The survey research literature is an invaluable resource for understanding and systematizing different error sources in the data collection and curation process that may surface as distribution shifts downstream. With their inherent focus on valid population inference, concepts such as under-coverage (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 identify and systematize deficits in the training data composition in relation to the target population. Error frameworks such as the Total Survey Error (Groves and Lyberg, 2010; Biemer, 2010) and its adaptions to online data sources (Sen et al., 2021) have been proposed to systematically trace errors along the data collection and processing pipeline that may accumulate in misrepresentation (and measurement error) in the eventual data available for model training.

Reflecting the growing use of alternative data sources for population inference, a parallel body of work proposes strategies for improving inference from data that was not collected using traditional sampling schemes (Cornesse et al., 2020; Yang and Kim, 2020). In this context, pseudo-weighting approaches (Elliott and Valliant, 2017; Valliant and 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 et al., 2017). As a recent example of cross-disciplinary work in this context, Kim et al. (2022) draw on the multicalibration framework from algorithmic fairness (Hebert-Johnson et al., 2018) to learn prediction functions that are universally adaptable to unknown deployment shifts and allow for downstream inference that is competitive with target-specific correction techniques.

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 (Kouw and Loog, 2018; Hedegaard et al., 2021). 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 et al., 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 et al., 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 (Shimodaira, 2000; Kouw and Loog, 2018). Estimating these weights makes it possible to express the target risk relative to the source distribution $R_{\mathcal{T}}(\mathcal{L}) = \mathbb{E}_{\mathcal{T}}[\mathcal{L}(X,Y)] = \mathbb{E}_{\mathcal{S}}[w(X,Y)\mathcal{L}(X,Y)]$ which then can be minimized (Fang et al., 2020). Various strategies can be used to estimate the importance weights, such as logistic regression (Bickel et al., 2009), kernel density estimation (Yu and Szepesvári, 2012), kernel mean matching (Quiñonero-Candela et al., 2009) and KL-divergence minimization (Sugiyama et al., 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 and Loog, 2018; Liu and 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 (Kouw and Loog, 2018; Subbaswamy et al., 2022; Duchi and Namkoong, 2021). For instance, Wen et al. (2014) focus on minimizing loss for the worst-case re-weighting of the source data within a predefined set of re-weighting functions.

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 $\tilde{Y}$. The label bias quantifies the difference between the true distribution of interest $p_T(Y|X)$, and the distribution $p_S(\tilde{Y}|X)$ 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 $\tilde{Y}$ (Gruber et al., 2023; Fang et al., 2020; Dai and Brown, 2020).

Various approaches have been devised to mitigate label noise and measurement error. For instance, Natarajan et al. (2013) propose an unbiased risk minimization strategy for handling class-conditional $p(\tilde{Y}|Y)$ noise. While such a simplified model of a proxy may be applicable in some settings, practitioners will likely encounter more complex scenarios, such as the measurement error depending on sensitive covariates (Obermeyer et al., 2019). In many situations, there may be the option to access multiple proxies of the true target of interest. For example, Boeschoten et al. (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 et al., 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 and Perdomo, 2022).

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 (Gruber et al., 2023; Bhatt et al., 2021; Sullivan, 2015; Hüllermeier and Waegeman, 2021) 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 (Vovk et al., 2022; Angelopoulos and Bates, 2021; Papadopoulos et al., 2002). A larger conformal set indicates higher uncertainty in the model's predictions. However, conformal prediction requires splitting the data into a training set and an additional holdout dataset, known as the calibration set. While this approach enhances model robustness, it comes at the expense of

reduced statistical efficiency due to the decreased amount of data available for training. Alternatively, full conformal prediction does not necessitate dividing the data but is usually computationally more demanding (Angelopoulos and Bates, 2021). Conformal predictions 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 (Tibshirani et al., 2019; Barber et al., 2023; Gibbs and Candes, 2021).

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 and Rodriguez (2023) 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 and 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. For example, if each individual intervention costs the same, a decision-maker can simply intervene on the top-k ranked individuals.

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 and 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 and Dennis, 1997), as stakeholders often struggle to determine how to exactly weigh different objectives (Hayes et al., 2022; Roijers et al., 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. Boutilier (2013) offers further insights into techniques for eliciting utility functions and preferences from decision-makers.

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 (Hayes et al., 2022; Deb, 2011). For example, Hertweck et al. (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 et al., 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 (Hort et al., 2023; Hardt et al., 2016b; Zafar et al., 2017; Corbett-Davies et al., 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 et al., 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 often involves counterfactual modeling. While non-causal risk predictions are often used as proxies for relevant counterfactual outcomes, they risk being significantly biased, making a more principled approach involving explicit causal modeling preferable. However, as outlined in Section 2.3, a common threat to the validity of causal models is confounding, requiring external assumptions and historical data on intervention assignment to address. This challenge requires careful case-by-case analysis by the model developer and may limit the possibility to make use of counterfactual estimates for decision-making in certain application contexts.

The potential outcomes framework (Rubin, 1974) is a prominent approach for framing causal questions. In a binary intervention scenario $\mathcal{T} = \{0, 1\}$, it denotes two potential outcomes $(Y_i(0), Y_i(1))$ for an individual i. These outcomes represent the two possible observable outcomes: no intervention $(T_i = 0)$ and an intervention $(T_i = 1)$. The individual treatment effect τ_i is then defined as the difference between the potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. Estimating potential outcomes and treatment effects from observational data $\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^n$ is challenging, as it is usually only possible to observe one outcome $Y_i = (1 - T_i)Y_i(0) + T_i Y_i(1)$ for each individual (Künzel et al., 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 $X = x$ (Vegetabile, 2021; Künzel et al., 2019).

To link the CATE with a statistical estimand, a set of untestable assumptions is required (Künzel et al., 2019; Caron et al., 2022; Johansson et al., 2022). Unconfoundedness $(Y(0), Y(1)) \perp\!\!\!\perp T|X$, requires that potential outcomes are conditionally independent of treatment assignment. Positivity guarantees nonzero propensity scores $0 < P(T = 1|X = x) < 1$ for all confounders $x \in \mathcal{X}$, meaning that treatment assignment is not fully deterministic. Finally, Stable Unit Treatment Value Assumption (SUTVA) assumes that the outcome of one individual is not affected by the interventions others received, and that there are no different versions of a specific treatment. Under these assumptions it becomes in principle possible to infer $\mu_t(x)$ and $\tau(x)$ from observational data (Caron et al., 2022). Ensuring these assumptions can be difficult, with unmeasured confounding posing a significant risk when aiming for valid counterfactual predictions. We refer to Appendix A for an overview of relevant ML-based CATE estimation methods.

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 and 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 (Van Geloven et al., 2020; Dickerman and Hernán, 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 (Coston et al., 2020; Chouldechova et al., 2018). 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 and 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 and 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 et al., 2023), allowing us to correctly prioritize individuals based on the estimated baseline risk $Y(0)$ alone (Fernández-Loría and 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; Van Geloven et al., 2020; Schulam and Saria, 2017; Prosperi 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 et al. (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 (Johansson et al., 2016). 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 and 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, Zhou et al. (2023) 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 et al., 2019), enabling unbiased predictions in a new deployment setting.

In a recent study, Guerdan et al. (2023b) 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 et al. (2023a) 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 et al., 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 tradeoffs 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 (Tu et al., 2021; Ai et al., 2022). 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 et al. (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, 2023). While perfect predictions could in theory lead to optimal decision-making, in practice it may be easier and more practical to estimate the allocation policy directly (Elmachtoub and Grigas, 2022; Fernández-Loría and 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 $\pi : \mathcal{X} \to [0, 1]$, directly mapping individual covariates X_i to the likelihood of intervening. Learning optimal assignment rules has been studied across different disciplines, such as statistical decision theory, economics and operations research (Manski, 2004; Kitagawa and Tetenov, 2018; Elmachtoub and Grigas, 2022). This paper specifically highlights recent literature focusing on learning optimal policies from past observational data using ML methods (Athey and Wager, 2021; Kallus, 2018; Hatamyar and Kreif, 2023; Luedtke and 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 $V(\pi) = \mathbb{E}[\pi(x)Y(0) + (1 - \pi(x))Y(1)]$ as the overall expected outcome if the policy were deployed (Athey and Wager, 2021). Finding the optimal policy corresponds to identifying the policy that maximizes the utility, i.e. $\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{V}(\pi)$. When estimating the policy value from observational data, we rely on the same assumptions as those used in CATE estimation to ensure identification, such as unconfoundedness. We refer to Appendix B 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. Firstly, 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 (Kallus, 2021; Athey and Wager, 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. Secondly, 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 and 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, as seen in (Athey and 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 et al. (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. Optimizing over a large neighborhood of distributions guarantees robustness against significant shifts, yet may substantially reduce the performance of the policy if the environment changes less than anticipated. 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 et al. (2022) incorporate DR methods, removing the need to assume that the historical assignment policy is known, which is often unavailable when relying on observational data.

However, domain experts may often encounter difficulties in precisely specifying the optimal neighborhood size (Hatt et al., 2022). Instead of ensuring robustness under arbitrary distribution shifts, it may be helpful to focus on specific types of shifts, potentially simplifying the integration of domain knowledge. For example, Hatt et al. (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 et al., 2022), for example as seen in Wang et al. (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 for individual covariates. Similar to CATE estimation, a critical challenge for conformal off-policy prediction lies in guaranteeing exchangeability of the data. Zhang et al. (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.

3.3.3 Multi-Objective Optimization

While there is significant research on multi-objective reinforcement learning (Hayes et al., 2022), the focus often lies on online policy learning and less attention is given to multi-objective policy learning from historical data. 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 and Xu, 2020; Sun, 2021). However, if the relevant constraint needs to be adjusted regularly, such approaches may lead to significant computational costs (Sun et al., 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.

Off-policy learning methods that optimize within a constrained class of policies (Athey and Wager, 2021; Kallus, 2021) have the advantage that secondary constraints can also be encoded through restricting the class of allowed policies. For example, Frauen et al. (2023) 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.

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

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.

Rehill and Biddle (2022) 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 Discussion

Making productive use of ML for public sector decision-making is a complicated task, requiring careful alignment of policy objectives and model development. We provided a overview of technical challenges that commonly arise when aligning policy objectives with ML models with the goal of enhancing decision-making in the public sector. We discussed the intricacies of acquiring representative training data and addressing issues like robustness under distribution shifts, dealing with proxy variables, and the influence of past decision-making. We also covered various challenges complicating the link between policy goal, model output and decision-making, including the difficulties of addressing multiple objectives and human involvement in decision-making.

Furthermore, we examined key target estimands and their alignment with decision-making

scenarios in the public sector: risk prediction, counterfactual prediction and policy learning. These modeling frameworks do not only add structure to decide and evaluate on how a given policy problem may be best connected or supported by ML approaches, but also point to context-specific solutions to the practical challenges outlined above. We highlighted relevant methodological advancements for causal estimation, handling distribution shifts, uncertainty quantification and multi-objective optimization that extend beyond the scope of the standard ML framework. We provide a distilled version of our presentation of different target estimands, 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. We recommend that policymakers and model developers collaboratively select the modeling approach most suited for their intended application by establishing a clear connection between the policy goal, allocation principle and ultimately the target estimand.

The intention behind this paper was to work toward 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.

Acknowledgment

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, 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 Felix Henninger, Patrick Schenk, Nanina Föhr 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.

References

Acharki, N., Lugo, R., Bertoncello, A., and 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., Zhang, Y., and 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., and van der Laan, M. (2023). Conformal Meta-learners for Predictive Inference of Individual Treatment Effects. *arXiv preprint arXiv:2308.14895*.

Alexopoulos, C., Lachana, Z., Androutsopoulou, A., Diamantopoulou, V., Charalabidis, Y., and Loutsaris, M. A. (2019). How Machine Learning is Changing E-Government. In *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance*, pages 354–363, New York. Association for Computing Machinery.

Algorithm Watch (2019). *Atlas of Automation – Automated Decision-Making and Participation in Germany*. AlgorithmWatch gGmbH. `https://atlas.algorithmwatch.org/en/`.

Algorithm Watch (2020). *Automating Society Report 2020*. AlgorithmWatch gGmbH & Bertelsmann Stiftung, Germany. `https://automatingsociety.algorithmwatch.org`.

Allhutter, D., Cech, F., Fischer, F., Grill, G., and Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, 3.

Amarasinghe, K., Rodolfa, K. T., Lamba, H., and Ghani, R. (2023). Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions. *Data & Policy*, 5:e5.

Angelopoulos, A. N. and Bates, S. (2021). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*.

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

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., and Kang, H. (2019a). 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., and 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., and Wager, S. (2019b). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

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

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

Bang, H. and 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.

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

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Belle, V. and Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4.

Bennett, A. and Kallus, N. (2019). Policy Evaluation with Latent Confounders via Optimal Balance. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bennett, A. and Kallus, N. (2020). Efficient Policy Learning from Surrogate-Loss Classification Reductions. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 788–798. PMLR.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., and 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, pages 401–413, New York, NY, USA. Association for Computing Machinery.

Bickel, S., Brückner, M., and 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., and Ho, D. (2022). Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1479–1503, Seoul Republic of Korea. ACM.

Boardman, A. E., Greenberg, D. H., Vining, A. R., and Weimer, D. L. (2018). *Cost-Benefit Analysis: Concepts and Practice*. Cambridge University Press, 5 edition.

Boeschoten, L., van Kesteren, E.-J., Bagheri, A., and Oberski, D. L. (2021). Achieving Fair Inference Using Error-Prone Outcomes. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(Special Issue on Artificial Intelligence, Paving the Way to the Future):9–15.

Boutilier, C. (2013). Computational Decision Support Regret-Based Models for Optimization and Preference Elicitation. In Zentall, T. R. and Crowley, P. H., editors, *Comparative Decision Making*, pages 423–453. Oxford University Press.

Caron, A., Baio, G., and 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., and Mao, X. (2023). Learning under Selective Labels with Data from Heterogeneous Decision-makers: An Instrumental Variable Approach. *arXiv preprint arXiv:2306.07566*.

Cheng, L. and Chouldechova, A. (2022). Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 134–148. PMLR.

Chugunova, M. and 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., and Bollens, J. (2023). Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics*, 80:102306.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and 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, page 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., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., and 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., and 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., and Chouldechova, A. (2020). Counterfactual Risk Assessments, Evaluation, and Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 582–593, New York, NY, USA. Association for Computing Machinery.

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

Curth, A. and 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*, pages 1810–1818. PMLR.

Curth, A. and 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*, pages 6623–6642. PMLR.

Dai, J. and Brown, S. M. (2020). Label Bias, Label Shift: Fair Machine Learning with Unreliable Labels. In *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*, volume 12.

Damen, J. A., Pajouheshnia, R., Heus, P., Moons, K. G. M., Reitsma, J. B., Scholten, R. J. P. M., Hooft, L., and Debray, T. P. A. (2019). Performance of the Framingham risk models and pooled cohort equations for predicting 10-Year risk of cardiovascular disease: A systematic review and meta-analysis. *BMC medicine*, 17(1):109.

Das, I. and 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., and Pal, D. (2010). Impossibility Theorems for Domain Adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings.

De-Arteaga, M., Fogliato, R., and 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, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Deb, K. (2011). Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction. In Wang, L., Ng, A. H. C., and Deb, K., editors, *Multi-Objective Evolutionary Optimisation for Product Design and Manufacturing*, pages 3–34. Springer, London.

Desiere, S. and 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. and 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., and 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., and 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. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*.

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

Duchi, J. C. and Namkoong, H. (2021). Learning Models with Uniform Performance via Distributionally Robust Optimization. *The Annals of Statistics*, 49(3):1378–1406.

Dudík, M., Langford, J., and Li, L. (2011). Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1097–1104, Madison, WI, USA. Omnipress.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., and 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*, 57:101994.

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

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

Enarsson, T., Enqvist, L., and 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., and Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper*, (20-54).

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 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*, 206.

Fang, T., Lu, N., Niu, G., and Sugiyama, M. (2020). Rethinking Importance Weighting for Deep Learning under Distribution Shift. In *Advances in Neural Information Processing Systems*, volume 33, pages 11996–12007. Curran Associates, Inc.

Fernández-Loría, C. and 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. and 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.

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

Fogliato, R., Chouldechova, A., and 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*, pages 2325–2336. PMLR.

Foster, D. J. and Syrgkanis, V. (2023). Orthogonal Statistical Learning. *The Annals of Statistics*, 51(3):879–908.

Frauen, D., Melnychuk, V., and Feuerriegel, S. (2023). Fair Off-Policy Learning from Observational Data. *arXiv preprint arXiv:2303.08516*.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761–767.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc.

Goddard, K., Roudsari, A., and 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. and 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., and Kauermann, G. (2023). Sources of Uncertainty in Machine Learning–A Statisticians' view. *arXiv preprint arXiv:2305.16703*.

Guerdan, L., Coston, A., Holstein, K., and Wu, Z. S. (2023a). Counterfactual Prediction Under Outcome Measurement Error. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1584–1598, Chicago IL USA. ACM.

Guerdan, L., Coston, A., Wu, Z. S., and Holstein, K. (2023b). 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, pages 688–704, New York, NY, USA. Association for Computing Machinery.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016a). Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 111–122, New York, NY, USA. Association for Computing Machinery.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016b). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Hatt, T., Tschernutter, D., and Feuerriegel, S. (2022). Generalizing off-policy learning under sample selection bias. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 769–779. PMLR.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., and 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., and Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948. PMLR.

Hedegaard, L., Sheikh-Omar, O. A., and 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., and 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., and Sarro, F. (2023). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*

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

Jeong, S. and Namkoong, H. (2020). Robust causal inference under covariate shift via worst-case subpopulation treatment effects. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2079–2084. PMLR.

Johansson, F. D., Shalit, U., Kallus, N., and 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., and Sontag, D. (2016). Learning Representations for Counterfactual Inference. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 3020–3029, New York, NY, USA. JMLR.org.

Kaiser, P., Kern, C., and 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 Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. 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., and Zhou, Z. (2022). Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10598–10632. PMLR.

Kang, J. D. Y. and 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. and Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press.

Kennedy, E. H. (2020a). Efficient Nonparametric Causal Inference with Missing Exposure Information. *The International Journal of Biostatistics*, 16(1).

Kennedy, E. H. (2020b). Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kerrigan, D., Hullman, J., and Bertini, E. (2021). A Survey of Domain Knowledge Elicitation in Applied Machine Learning. *Multimodal Technologies and Interaction*, 5(12):73.

Kim, K. and Zubizarreta, J. R. (2023). Fair and robust estimation of heterogeneous treatment effects for policy learning. In *International Conference on Machine Learning*, pages 16997–17014. PMLR.

Kim, M. P., Ghorbani, A., and 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, pages 247–254, New York, NY, USA. Association for Computing Machinery.

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., and Reingold, O. (2022). Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119.

Kim, M. P. and Perdomo, J. C. (2022). Making decisions under outcome performativity. *arXiv preprint arXiv:2210.01745*.

Kitagawa, T. and 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., and Mullainathan, S. (2018). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293.

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

Körtner, J. and Bach, R. (2023). Inequality-Averse Outcome-Based Matching. Preprint, OSF Preprints. `https://osf.io/yrn4d/`.

Körtner, J. and Bonoli, G. (2023). Predictive algorithms in the delivery of public employment services. In *Handbook of Labour Market Policy in Advanced Democracies*, pages 387–398. Edward Elgar Publishing.

Kouw, W. M. and Loog, M. (2018). An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*.

Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and 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., and 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*, 7:883999.

Kuzmanovic, M., Frauen, D., Hatt, T., and Feuerriegel, S. (2024). Causal machine learning for cost-effective allocation of development aid. *arXiv preprint arXiv:2401.16986*.

Kuzmanovic, M., Hatt, T., and Feuerriegel, S. (2023). Estimating Conditional Average Treatment Effects with Missing Treatment Information. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 746–766. PMLR.

Lakkaraju, H. and 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, pages 79–85, New York, NY, USA. Association for Computing Machinery.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and 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*, pages 275–284, Halifax NS Canada. ACM.

Laux, J., Wachter, S., and 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., and 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. and 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., and 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., and 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., and 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., and Smola, A. (2018). Detecting and Correcting for Label Shift with Black Box Predictors. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130. PMLR.

Liu, A. and Ziebart, B. (2014). Robust Classification Under Sample Selection Bias. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Luedtke, A. R. and 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. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, 86(3):532–565.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

Mayer, A.-S., Strich, F., and Fiedler, M. (2020). Unintended consequences of introducing ai systems for decision making. *MIS Quarterly Executive*, 19(4).

McFowland III, E., Gangarapu, S., Bapna, R., and 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., and 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., and 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*, pages 386–400.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition. `https://christophm.github.io/interpretable-ml-book/`.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and 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., and 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., and Tewari, A. (2013). Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Nie, X. and Wager, S. (2020). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*, 108(2):299–319.

Nourani, M., Kabir, S., Mohseni, S., and 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., and Sendhil Mullainathan (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., and Wu, Z. S. (2019). Orthogonal Random Forest for Causal Inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4932–4941. PMLR.

Pagan, N., Baumann, J., Elokda, E., De Pasquale, G., Bolognani, S., and Hannák, A. (2023). A classification of feedback loops and their relation to biases in automated decision-making systems. *arXiv preprint arXiv:2305.06055*.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, ECML'02, pages 345–356, Berlin, Heidelberg. Springer-Verlag.

Papalexopoulos, T. P., Bertsimas, D., Cohen, I. G., Goff, R. R., Stewart, D. E., and 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. and Barocas, S. (2019). Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48, Atlanta GA USA. ACM.

Pencheva, I., Esteve, M., and 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., and Hardt, M. (2020). Performative Prediction. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR.

Perdomo, J. C. (2023). The relative value of prediction in algorithmic decision making. *arXiv preprint arXiv:2312.08511*.

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

Pfisterer, F., Coors, S., Thomas, J., and Bischl, B. (2019). Multi-objective automatic machine learning with autoxgboostmc. *arXiv preprint arXiv:1908.10796*.

Post, R., van den Heuvel, I., Petkovic, M., and van den Heuvel, E. (2022). Flexible machine learning estimation of conditional average treatment effects: A blessing and a curse. *arXiv preprint arXiv:2210.16547*.

Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgenson, E., Mansour, R., and 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*, pages 2039–2047, Sydney NSW Australia. ACM.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–52, New York, NY, USA. Association for Computing Machinery.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*, pages 131–160. MIT Press.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2008). *Dataset Shift in Machine Learning*. MIT Press.

Rambachan, A., Coston, A., and Kennedy, E. (2022). Counterfactual Risk Assessments under Unmeasured Confounding. *arXiv preprint arXiv:2212.09844*.

Rehill, P. and Biddle, N. (2022). Policy learning for many outcomes of interest: Combining optimal policy trees with multi-objective bayesian optimisation. *arXiv preprint arXiv:2212.06312*.

Robins, J. M., Rotnitzky, A., and 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., and 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., and Zhong, C. (2022). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys*, 16(none):1–85.

Salditt, M., Eckes, T., and 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*, pages 1–24.

Schulam, P. and Saria, S. (2017). Reliable Decision Support using Counterfactual Models. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1):399–422.

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.

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., and Blanchet, J. (2020). Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8884–8894. PMLR.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2023). Distributionally Robust Batch Contextual Bandits. *Management Science*.

Singh, K., Valley, T. S., Tang, S., Li, B. Y., Kamran, F., Sjoding, M. W., Wiens, J., Otles, E., Donnelly, J. P., Wei, M. Y., McBride, J. P., Cao, J., Penoza, C., Ayanian, J. Z., and 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. and Rodriguez, M. G. (2023). Designing Decision Support Systems Using Counterfactual Prediction Sets. *arXiv preprint arXiv:2306.03928*.

Subbaswamy, A., Chen, B., and 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., and Kawanabe, M. (2007). Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Sullivan, T. (2015). *Introduction to Uncertainty Quantification*, volume 63 of *Texts in Applied Mathematics*. Springer International Publishing, Cham.

Sun, H., Munro, E., Kalashnov, G., Du, S., and Wager, S. (2024). Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066*.

Sun, L. (2021). Empirical welfare maximization with constraints. *arXiv preprint arXiv:2103.15298*, 2.

Sun, T. Q. and 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. and Joachims, T. (2015). Counterfactual Risk Minimization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 939–941, Florence Italy. ACM.

Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., and Doucet, A. (2022). Conformal off-policy prediction in contextual bandits. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31512–31524. Curran Associates, Inc.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal Prediction Under Covariate Shift. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tu, Y., Basu, K., DiCiccio, C., Bansal, R., Nandy, P., Jaikumar, P., and Chatterjee, S. (2021). Personalized treatment selection using causal heterogeneity. In *Proceedings of the Web Conference 2021*, WWW '21, page 1574–1585, New York, NY, USA. Association for Computing Machinery.

Vaithianathan, R., Benavides-Prado, D., Dalton, E., Chouldechova, A., and 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., and Benavides-Prado, D. (2019). Allegheny family screening tool: Methodology, version 2. *Center for Social Data Analytics*, pages 1–22.

Valliant, R. and 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., Groenwold, R. H. H., Van Houwelingen, H. C., Putter, H., and Le Cessie, S. (2020). Prediction meets causal inference: The role of treatment in clinical prediction models. *European Journal of Epidemiology*, 35(7):619–630.

van Noordt, C. and 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*, page 101714.

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., and Zou, J. Y. (2022). Uncalibrated Models Can Improve Human-AI Collaboration. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4004–4016. Curran Associates, Inc.

Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Springer International Publishing, Cham.

Wager, S. and 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., and 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, page 626, New York, NY, USA. Association for Computing Machinery.

Wang, Y.-X., Agarwal, A., and 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, page 3589–3597. JMLR.org.

Watson-Daniels, J., Barocas, S., Hofman, J. M., and 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, page 297–311, New York, NY, USA. Association for Computing Machinery.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., and Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994.

Wen, J., Yu, C.-N., and 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*, pages 631–639. PMLR.

Wirtz, B. W., Weyerer, J. C., and 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., Weintraub, W. S., Moran, A. E., and Shen, J. (2022). Estimating the optimal individualized treatment rule from a cost-effectiveness perspective. *Biometrics*, 78(1):337–351.

Yang, S. and 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., and 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., and Tao, D. (2020). Label-Noise Robust Domain Adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10913–10924. PMLR.

Yu, Y.-L. and 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, pages 1147–1154, Madison, WI, USA. Omnipress.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 962–970. PMLR.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012). Estimating Optimal Treatment Regimes from a Classification Perspective. *Stat*, 1(1):103–114.

Zhang, Y., Shi, C., and Luo, S. (2023). Conformal off-policy prediction. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2751–2768. PMLR.

Zhou, A., Kern, C., and Kim, M. (2023). Robust Conditional Average Treatment Effect Estimation via Multi-Accurate Learning. `https://simons.berkeley.edu/talks/christoph-kern-lmu-munich-2023-04-26`.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Zuiderwijk, A., Chen, Y.-C., and 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):101577.

A CATE Estimation Methods

In recent years, several CATE estimation strategies have emerged, many of which employ non-parametric ML regression models to estimate the relationship between covariates, outcome and intervention (Lechner, 2023; Caron et al., 2022). 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 (Künzel et al., 2019; Caron et al., 2022; Curth and van der Schaar, 2021).

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 and 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 (Johansson et al., 2022; Nie and Wager, 2020; Caron et al., 2022; Künzel et al., 2019).

Alternative methods directly estimate the CATE function by first constructing pseudo-outcomes of the treatment effects (Künzel et al., 2019; Caron et al., 2022; Curth and van der Schaar, 2021). 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 and van der Schaar, 2021). The Doubly-Robust learner (DR-learner) employs pseudo-outcomes constructed from both the conditional outcomes $\hat{\mu}_t(x)$ and the propensity scores $\hat{\pi}(x)$ (Kennedy, 2020b). DR estimators have a long history in causal inference and missing data imputation (Kang and Schafer, 2007; Bang and Robins, 2005; Robins et al., 1994; Funk et al., 2011) 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 (Nie and Wager, 2020; Foster and Syrgkanis, 2023) 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 et al. (2019b); Wager and Athey (2018), resemble the R-learner (Caron et al., 2022; Post et al., 2022; Oprescu et al., 2019).

A large body of literature analyzes the asymptotic and finite sample properties of different CATE learners (Salditt et al., 2023; Curth and van der Schaar, 2021; Künzel et al., 2019; Caron et al., 2022; Kennedy, 2020b). 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.

B 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 and 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 $\hat{V}$ (Qian and Murphy, 2011; Bennett and Kallus, 2020). 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 and Wager, 2021; Zhang et al., 2012; Dudík et al., 2011). Several approaches have been proposed to relax the assumptions for causal identification, for example methods that address learning policies under unmeasured confounding (Bennett and Kallus, 2019) or handle situations with limited overlap (Kallus, 2021).