Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making

Luke Guerdan
Carnegie Mellon University
Pittsburgh, USA
lguerdan@cs.cmu.edu

Amanda Coston
Carnegie Mellon University
Pittsburgh, USA

Zhiwei Steven Wu
Carnegie Mellon University
Pittsburgh, USA

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, USA

ABSTRACT

A growing literature on human-AI decision-making investigates strategies for combining human judgment with statistical models to improve decision-making. Research in this area often evaluates proposed improvements to models, interfaces, or workflows by demonstrating improved predictive performance on "ground truth" labels. However, this practice overlooks a key difference between human judgments and model predictions. Whereas humans reason about broader phenomena of interest in a decision – including latent constructs that are not directly observable, such as disease status, the "toxicity" of online comments, or future "job performance" – predictive models target *proxy* labels that are readily available in existing datasets. Predictive models' reliance on simplistic proxies makes them vulnerable to various sources of statistical bias. In this paper, we identify five sources of *target variable bias* that can impact the validity of proxy labels in human-AI decision-making tasks. We develop a causal framework to disentangle the relationship between each bias and clarify which are of concern in specific human-AI decision-making tasks. We demonstrate how our framework can be used to articulate implicit assumptions made in prior modeling work, and we recommend evaluation strategies for verifying whether these assumptions hold in practice. We then leverage our framework to re-examine the designs of prior human subjects experiments that investigate human-AI decision-making, finding that only a small fraction of studies examine factors related to target variable bias. We conclude by discussing opportunities to better address target variable bias in future research.

KEYWORDS

algorithmic decision support, measurement, validity, causal diagrams, label bias, human-AI decision-making

1 INTRODUCTION

A growing body of research combines predictive machine learning models with human judgment to improve decision-making processes. This work often pursues the goal of *complementary performance*: configurations of humans and models that yield higher-quality decisions than either would make in isolation. In the machine learning community, model-level improvements have been proposed to address gaps in human judgment (e.g., [46, 62, 91, 96]). In the human-computer interaction community, behavioral interventions have been developed to help *humans* better incorporate model outputs into their decision-making (e.g., [4, 9]). However, current evaluations of complementary performance make the **key assumption that the label targeted by a predictive model adequately reflects the goals of human decision-makers**. Decision quality is frequently operationalized by predictive performance, measured via accuracy, AU-ROC, or similar statistical measures computed with respect to *"ground truth"* labels that are readily available in existing data. The relative performance of human judgment, model predictions, and hybrid combinations of the two are then ranked according to these metrics. Yet such comparisons of human and model performance are only valid insofar as the *"ground truth"* label targeted by the model reflects the underlying objectives of human decision-makers.

In real-world human-AI decision-making settings, labels are often imperfect proxies for the target outcomes considered by decision-makers. While making decisions, humans frequently consider latent constructs such as the "toxicity" of an online comment, "cardiovascular disease risk" of a patient, or future "job performance" of a candidate. Because observed labels (e.g., toxicity annotations, diagnostic test results, and performance reviews) serve as indirect measurements of these phenomena [49], they can be subject to *measurement error*. Additionally, humans often select among multiple possible actions (e.g., medical treatments, social welfare interventions) in hopes of improving a downstream outcome of interest. Because an outcome is only observed for the selected action, labeled data does not contain the counterfactual outcome that *would* occur had a different option been chosen instead. This introduces a set of additional challenges, including selective labels [60], intervention effects [86], and selection bias, which interact with measurement error in nuanced ways depending on the nature of the decision-support task. We refer to this collection of challenges – which can be characterized as sources of statistical bias impacting labels – as *target variable bias*.[1]

Target variable bias has been widely documented in real-world deployments of algorithmic systems. Predictive models impacted by target variable bias have contributed to unwarranted firing of teachers [14], perpetuated historical disparities in access to medical resources [71], and raised concerns among social workers investigating allegations of child abuse and neglect [16, 52]. Surprisingly, existing modeling efforts and human subjects experiments in the

[1]The term Target Variable Bias was introduced in [17, 31]. We use this as an umbrella term describing sources of statistical bias known to impact proxy labels in decision support tasks.

human-AI decision-making literature have largely overlooked this challenge. Left unaddressed, this disconnect could undermine the ultimate goal of human-AI decision-making research: to develop algorithmic systems that meaningfully improve decision-making in real-world contexts. In this paper, we aim to bridge the gap between challenges encountered in real-world deployments of predictive models and current human-AI decision-making research practices by (i) raising awareness of target variable bias, (ii) identifying blind spots in previously published modeling approaches and human subjects experiments, and (iii) providing guidelines for improved research practices going forward.

Based on an examination of numerous real-world deployments of predictive models, we construct a causal framework that captures possible data generating processes in human-AI decision-making. Our framework (1) distills high-level structural differences between various real-world decision support tasks and (2) disentangles which sources of bias are of concern in each setting. We map our framework to existing terminology and methods used to characterize target variable modeling assumptions across multiple disciplines. Using our framework, we then identify strategies for better-addressing target variable bias through two lines of human-AI decision-making research:

- **Model development**. We develop a *measurement* and *prediction* decomposition that articulates target variable modeling assumptions. We use our decomposition to create a taxonomy of model-level improvements proposed in previous literature. We also propose a set of recommended measurement model evaluation strategies.
- **Experimental human subjects studies**. Using our framework, we re-examine the design of prior human subjects experiments studying human-AI decision-making. Our analysis identifies systematic blind spots in our current understanding of human-AI decision-making due to target variable bias.

2 RELATED WORK

We begin by introducing the body of human-AI decision-making research our framework is designed to inform. We then summarize modeling challenges and broader validity concerns that draw current research practices (i.e., modeling assumptions, experimental study designs, and measures of decision quality) into question.

2.1 Human-AI decision-making

Recent machine learning research proposes techniques to complement the limitations of human judgement. Drawing from a long line of work showing that actuarial risk assessments can outperform expert judgement in many prediction tasks [22, 40], methods have been proposed that learn to complement humans by adaptively routing decision instances [36, 62], leveraging heterogeneity in human and machine decision performance [15, 29, 91, 96], leveraging consistency in expert decisions [24], and adapting to [46] and training [66] human mental representations of model outputs. Yet these techniques operate on a set of *simplifying assumptions* about the world, which may or may not hold in a given deployment context. We provide a framework for articulating modeling assumptions, and show that many assumptions made by prior work involving proxy labels are unlikely to hold in practice. Recent research has also studied opportunities for human-AI complementary in algorithm-assisted *human* decision-making [9, 13, 35, 38, 58, 59]. This work investigates the potential for tools such as training protocols [12, 13, 58], explanations [10, 61], and other behavioral interventions [9, 35], to improve how humans make use of model outputs. While many online experimental studies have been focused on interventions to improve predictive performance, little work to date has experimentally studied other key factors that are present in real-world deployment contexts, such as asymmetric access to information [45, 47], measurement error [37], and omitted payoffs [39].

2.2 Modeling challenges in algorithmic decision support

Prior work has surfaced a litany of challenges impacting predictive models designed for algorithmic decision support (ADS), including unobservables [53], selective labels [60], selection bias [23, 87], and intervention effects [20]. Additional work has examined the quality of proxy labels in decision support tasks. For example, Obermeyer et al. [71] surfaced *"label choice bias"*, in which racial disparities in access to health resources were introduced by poor label selection decisions. *"Omitted payoffs bias"* describes factors of interest to humans that are incompletely reflected by predictive models targeting available labels [14, 24, 53]. While this bias describes challenges specific to prediction (e.g., model unobservables, measurement error [53]), this term also applies when humans care about a broader set of decision-making factors beyond predictive risk [14, 38]. In this work, we use the lens of *measurement* to examine systematic differences between target outcomes of interest to humans and proxy labels observed in data [49]. In adopting this lens, we draw upon a rich set of existing knowledge and methodologies from adjacent disciplines (e.g., psychology, political science, sociology) designed to evaluate how latent phenomena of interest to humans are quantified in data [83].

2.3 Measurement and validity in algorithmic systems

Recent work has raised broader concerns regarding whether algorithmic systems successfully achieve their purported function [5, 49**?**]. Synthesizing concepts from measurement theory in the quantitative social sciences, Jacobs and Wallach [49] argue that *"algorithmic fairness"* is a latent construct that is imperfectly operationalized by statistical fairness measures. Bao et al. [5] examine statistical biases present in criminal justice datasets (e.g., ProPublica's COMPAS Dataset [3]) used in fairness benchmarks of algorithmic Risk Assessment Instruments (RAIs). This analysis identified several biases in the *outcome variable* Y predicted by models, which we further characterize in this work. **?**] also highlight a host of validity concerns impacting RAIs, including many discussed in Section 2.2. Recent work has also surfaced validity issues in content moderation [37] and recommender systems [65, 89]. Despite this growing awareness, we currently lack a holistic understanding of validity threats to prediction targets in human-AI decision-making. Addressing this gap is critical for preventing algorithmic harms in real-world deployment contexts. Therefore, in this work, we offer the first holistic examination of how measurement error, unobservables, selection bias, and treatment effects interact to impact target variable validity in real-world ADS deployments.

3 FRAMEWORK

We now describe our framework scope (Section 3.1) and development process (Section 3.2) before introducing our causal diagram (Section 3.3). We then use our framework to identify structural differences between different ADS tasks (Section 3.4) and disentangle sources of target variable bias (Section 3.5).

3.1 Scope

Our framework applies to settings in which a supervised learning model is introduced to augment human decision making by predicting (i) a future event (e.g., medical [71], criminal justice [28], child welfare [17], or real estate [47, 81] related outcomes); (ii) a subjective human annotation (e.g., perceived content toxicity [37]); or (iii) factual information (e.g., food nutrition [9]). In these settings, model predictions are combined with human decision-making, either by showing predictions to a human (i.e., algorithm-in-the-loop [38]), who makes the final decision, or via a hybrid flow of agency (e.g., deferral-based learning [62], learning with bandit feedback [36]). Given our focus on prediction-based decision tasks, we do not directly examine decision-support settings involving unsupervised learning (e.g., clustering), tasks relying upon *generative models* (e.g., text or image generation), or sequential settings with time dependency (e.g., reinforcement learning) in this work.

3.2 Framework development

Understanding which statistical biases are of concern in a given ADS task requires *examining the historical data generating process* that gave rise to the model training dataset. Causal diagrams, which are graphs that show causal relationships between nodes via connected edges [75], are tools specifically designed for this purpose. A causal diagram shows each variable being modeled as a discrete *node*. Causal connections between nodes are shown via *edges*. If the direction of a causal pathway is known, this is shown via a directed arrow from the parent to child node. An undirected edge is used to connect nodes when the causal direction is unknown or varies [75]. Our framework constructs a causal diagram to examine challenges impacting the labels available in data. Therefore, we specifically consider variables (i.e., *nodes*) and relationships (i.e., *edges*) that directly relate to the target variable; we *abstract away* other important factors, such as the training [12, 13, 58], decision-making process [39], and workflow [38] of the human decision-makers using the predictive model. While prior work has examined these factors in detail [12, 13, 38, 39, 58], our framework foregrounds factors most salient for target variable bias. In Section 5.3, we outline how our approach can be extended to systematically examine a broader set of components beyond target variables in ADS research.

Our causal diagram was developed and refined through an iterative series of discussions among the authors and external researchers spanning a range of disciplines. Based on a review of real-world case studies (see Appendix A), we synthesized candidate causal diagrams that could adequately characterize the target variable of interest across settings, and then stress-tested these diagrams by attempting to identify counterexamples. Through our discussions with external researchers, we also cross-referenced our framework with existing terminology and methods developed in adjacent disciplines, such as medical diagnostic testing, educational assessment, behavioral health, and statistics.

3.3 Causal diagram

3.3.1 Diagram structure. Figure 1 shows our proposed causal diagram, which represents **a space of directed acyclic graphs (DAGs)** describing different possible relationships between predictors, decisions, target variables, and their proxies in human-AI decision-making settings.

Predictors. X describes *covariates* used to generate model predictions. Covariates are often drawn from administrative data sources (e.g., medical records, lending history) available to an organization for model development. In ADS settings, humans can also make use of *unobserved contextual information* Z while making decisions. For example, a physician might consider real-time medical test results (e.g., electrocardiograms [68]) unavailable to a model, while a social worker might weigh contextual factors described via phone calls while deciding whether to recommend investigation of child maltreatment allegations [52]. In some cases, human decision-makers can also be unaware of a subset of covariates (e.g., due to organizational policy or prohibitively large datasets) [47]. Figure 1 refers to X and Z as *model observables* and *model unobservables*, respectively, based on whether the predictors are available to a model.

Decisions. The blue shaded box in Figure 1 shows the joint human-algorithm decision D. We decompose this node into separate variables for human decisions (D_H) and algorithm predictions (D_A). Prior to deployment of an algorithm, decisions result solely from human judgement D_H. In some cases, decisions post-deployment result from humans incorporating predictions into their decision-making (i.e., algorithm-in-the-loop [38]). In other cases, the joint decisions result from a learned combination of D_H and D_A [15, 29, 36, 46, 62, 91, 96].

Target variables. The node Y^* describes the unobserved target outcome of interest to human decision-makers. For example a model might be introduced to weigh risk of unobserved constructs such as "medical need" , "recidivism", "creditworthiness", or "job performance." Y describes the *observed proxy* that is targeted by a model in place of Y^*. For example, a model might predict "cost of medical care" [71], "re-arrest" [31], "loan default", or "supervisor performance reviews" in place of the targets listed previously. The grey box in Figure 1 represents a *measurement model* mapping the unobserved construct to the observed proxy targeted by a predictive model (see Section 3.5.1).

Edges. *We now describe the space of possible relationships across ADS tasks, before narrowing in on several special cases in Section 3.4.* Across ADS tasks, covariates and unobservables both contribute to human decisions D_H, while algorithmic predictions D_A are only influenced by covariates X. For example, a physician might make use of medical records (X) and real-time test results (Z), while an algorithm only has access to medical records (X). We show these relationships via directed arrows $X \rightarrow D$ and $Z \rightarrow D_H$. Decisions D can also influence the target *and* proxy outcomes Y^* and Y (e.g., as when medical treatments impact health status (Y^*) and cost

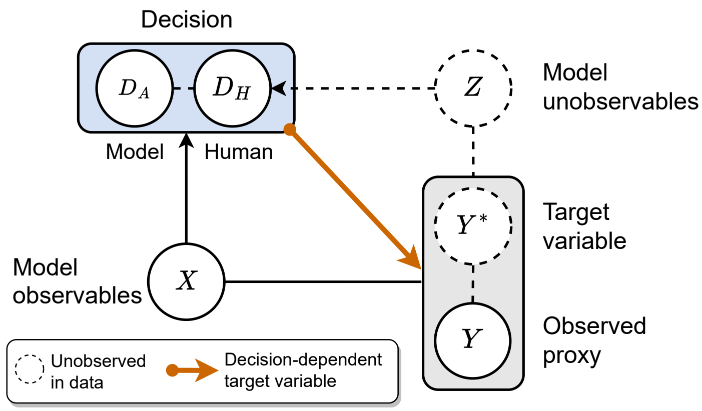

Figure 1: Our causal diagram represents a space of causal graphs, spanning different possible relationships between predictors, decisions, target variables, and their proxies in human-AI decision-making settings. Edges with directionality that can vary across different real-world human-AI decision settings are indicated via undirected edges. Observed variables are shown with solid lines, while unobserved variables are shown in dotted lines. An arrow pointing to a shaded box is shorthand for separate arrows pointing from the source to nodes contained within the box.

(Y)). The direction of the causal relationship between covariates (X), unobservables (Z), and prediction targets (Y and Y^*) can vary across specific ADS domains. We convey ambiguity in the direction of causation via undirected edges and describe this nuance below.

3.4 Special cases of our causal diagram

The generalized causal diagram we propose in Figure 1 can be used to articulate key structural differences between decision support tasks studied in human-AI decision-making literature. Identifying the task-specific diagram applicable to a given decision support setting is critical for (1) identifying relevant sources of target variable bias (Section 3.5), (2) articulating modeling assumptions (Section 4), and (3) designing ecologically valid studies (Section 5).

3.4.1 Decision-dependent target variables. A decision support task contains decision-dependent target variables when *the decision informed by an algorithm also impacts the downstream outcomes Y and Y^**. Real-world deployments of risk assessments often contain decision-dependent targets. For example, re-arrest is only observed among defendants released on bail [53]. Child welfare screening decisions influence the underlying risk of child maltreatment, in addition to observed proxies (i.e., placement in foster care) [20]. More generally, real-world decisions informed by algorithms often constitute *risk mitigating interventions* (e.g., medical treatments, educational programs) or *opportunities* (e.g., loans, new candidate hires) that change the likelihood of the target outcome (e.g., disease prognosis, educational attainment). Settings with decision-dependent outcomes contain the orange arrow from D to Y and Y^* shown in Figure 1.

3.4.2 Decision-independent target variables. In contrast, the target variable is *not* influenced by the proposed decision in *decision-independent* ADS tasks. While this setting is uncommon in real world deployments of algorithmic systems, human-AI decision-making studies often adopt tasks with decision-independent target variables. For instance, studies have examined models that predict factual content (e.g., food nutrition [9]) and perceptual information (e.g., counts of objects [73], geometric shapes [97]). In these settings, the prediction target (i.e., food nutrition, geometric shape) is *not* influenced by the prediction made by a human and/or model. Therefore, settings with a decision-independent target variable *do not contain* the arrow from D to Y and Y^* in Figure 1.

3.4.3 Subjective annotation. In some contexts, the target variable in question is a latent construct (e.g., toxicity or emotional state) operationalized by *subjective human annotations*. Here, human annotators serve as a *measurement model* linking the target construct (e.g., toxicity, hate speech) to a proxy (i.e., individual annotation) [49]. Because humans can disagree considerably on subjective tasks [37], labels elicited from specific individuals (Y) are an imperfect proxy for the broader construct of interest (Y^*). In subjective human annotation tasks, D constitutes decisions made at *runtime* (i.e., after deployment of system predicting the target construct). For example, these decisions can involve removal of content flagged as "toxic" or reversal of "damaging" Wikipedia edits [41, 56].

3.4.4 Considerations of reverse causality. Traditional notions of prediction hold that covariates (X) and unobservables (Z) contribute to downstream outcomes (Y) and (Y^*) via a domain-specific causal pathway [7]. In our diagram, this flow of information would be communicated via edges from X to (Y, Y^*), and from Z to (Y, Y^*). However, in some cases, the causal pathway can be *reversed* [43]. For example, this is possible if a patient's unobserved disease status (Y^*) contributes to their medical history (X) and real-time test results (Z). This ambiguity can also exist for target and proxy outcomes. Therefore, we communicate this domain-specific variation in causal directionality via undirected edges in Figure 1. [2]

[2]In order for the causal diagram to remain valid (i.e., a directed *acyclic* graph), one of the edges connecting nodes must remain disconnected in these settings (i.e., when target variables are decision-dependent and $Y^* \rightarrow X$, $Y \rightarrow X$, $Y^* \rightarrow Z$, or $Y \rightarrow Z$). While this requirement is consistent with the scope of our framework, which considers non-sequential settings, feedback loops are an important factor to consider in sequential settings [30].

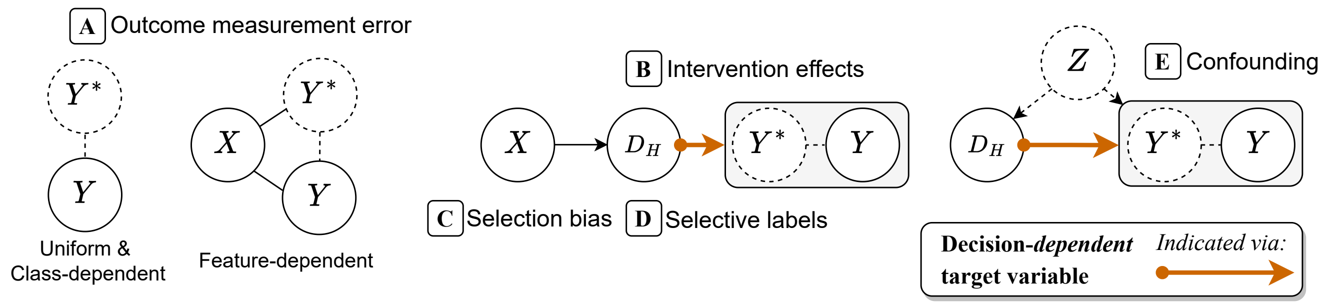

Figure 2: Sub-graphs of the diagram in Figure 1 introducing statistical biases that impact the target variable Y^*. *Outcome measurement error* (A) can occur in settings with both decision-dependent and independent target variables. In decision-dependent settings, *intervention effects* (B), *selection bias* (C), *selective labels* (D), and *confounding* (E) are also of concern.

3.5 Sources of target variable bias

We now use our causal diagram (Figure 1) to examine five sources of target variable bias that can threaten the validity of prediction targets. Critically, *our analysis shows that tasks with decision-dependent outcome variables introduce a broader set of challenges than tasks with decision-independent outcomes.* This suggests that greater care may be required during model development (Section 4) and experiment design (Section 5) in decision-dependent outcome tasks.

3.5.1 Outcome measurement error. Human experts and policy-makers often make decisions involving unobserved, latent constructs such as *"recidivism risk"* and *"job performance."* These latent constructs are not directly observable in the world, but can be operationalized via a *measurement model* [42, 49]. Adopting a label observed in data as a proxy for an unobserved latent construct serves as a *de facto* measurement model. For instance, in criminal justice settings, defendant re-arrest is commonly adopted as a proxy for recidivism risk [5, 33], while in commercial hiring settings, manager reviews are frequently adopted as a proxy for future job performance. Outcome measurement error (Figure 2.A) occurs when there is a systematic difference between the target variable of interest to experts and policy-makers (Y^*) and its operationalization by a proxy (Y). This challenge has been extensively documented in judicial [11, 31], child welfare [16, 52], and hiring [14] ADS domains.

Critically, **proxy labels impacted by measurement error offer an incomplete reflection of the actual goals of human decision-makers, and therefore serve as an incomplete measure of human-AI decision quality**. Therefore, before adopting a proxy to evaluate human-AI decision-making, it is critical to assess whether it serves as a satisfactory approximation of the target variable of interest. Measurement theory in the quantitative social sciences provides tools to conduct this assessment by weighing the *construct validity* and *reliability* of observed labels [42, 49] (see Section 4.3.1). In practice, measurement error in proxies is often studied via *measurement error models.* These models make *assumptions on the relationship* between the target outcome (Y^*) and its proxy (Y) (see Appendix B).

3.5.2 Intervention effects. In many ADS tasks, decisions serve as *risk mitigating interventions* intended to improve the chances of a favorable policy-relevant outcome [6, 20, 55]. As a result, past human decisions D_H influence the probability of the target outcome Y^* and its proxy Y (Figure 2.B). However, many existing predictive techniques mistakenly assume that decisions D and outcomes Y, Y^* are statistically independent [6, 20, 55]. This practice can be traced back to formulation of ADS as a prediction-policy problem [54], in which models are trained to maximize predictive performance with respect to observed outcomes without considering causal effects from D to Y and Y^*. Yet, we argue that accounting for the causal connection between decisions and outcomes is of central interest in many ADS tasks. For instance, consider two distinct policy problems that arise in tasks with decision-dependent target variables:

- **Selective Intervention (SI):** In this policy setting, organizations provide resources to individuals who are at *high baseline risk under no intervention.* For example, developers of the Allegheny Family Screening Tool (AFST) introduced the tool with the goal of assessing *"latent risk"* of maltreatment prior to county child welfare interventions [92]. Similarly, predictive models have been introduced in educational settings to identify students at-risk of failing given no tutoring resources [88]. This task requires causal inference because it involves inferring what *would occur* if an individual does not receive the proposed intervention.
- **Selective Opportunity (SO):** In this policy setting, an organization grants an opportunity (e.g., a new loan, or pre-trial release on bail) to decision subjects while trying to minimize risk of an adverse outcome (e.g., loan default, recidivism) given an individual receives the opportunity. This prediction task requires causal inference because it involves predicting what would occur under the *hypothetical scenario* that an individual receives the opportunity under consideration.

Naively structuring an ADS task as a prediction policy problem in SI and SO settings can lead to misleading assessments of model performance. For example, Coston et al. [20] demonstrate that predicting observational modeling and evaluation in SI settings systematically underestimates the risk for high-risk individuals who would respond most favorably to the intervention. The underlying modeling and evaluation challenge introduced by intervention effects stems from the fact that downstream outcomes are only observed under one of the possible decisions [50, 76, 78]. This challenge is closely related to an additional source of target variable bias: *selective labels* [60].

3.5.3 Selective labels. Another challenge introduced by the edge from D to Y, Y^* in Figure 2 is *selective labels*. This bias has been widely discussed in connection to pre-trial risk assessments, where recidivism-related proxy outcomes (e.g., re-arrest) are only observed among defendants released on bail [5, 33, 53, 60]. Selective labels also occur in child welfare settings, in which some outcomes (e.g., placement in foster care) are only observed among cases screened-in for investigation [23]. Selective labels maps directly to *selective intervention* and *selective opportunity* policy problems because we never observe how an individual *would have* benefited from a missed opportunity (SO), or how an intervention *would have* impacted an individual who historically received no additional resources. Selective labels pose the greatest challenge when *selection bias* was also present in the historical data generating process.

3.5.4 Selection bias. This bias, which occurs when features (X) or unobservables (Z) influenced past decisions (D) (Figure 2.C), complicates selective labels and intervention effects. Because a previous decision-making policy may have been more likely to intervene (SI) or grant opportunities (SO) to some sub-populations, these groups may be systematically over- or under- represented in historical outcome data. As a result, ADS models trained on historical data will not perform equally well on all sub-populations during deployment [8]. This effect has been well-documented in recidivism prediction settings, in which models predicting re-arrest outcomes have worse performance among sub-populations historically denied bail [51]. While re-weighting techniques can correct for selection bias [90], these approaches typically assume that no model unobservables Z are present, which can be unreasonable in many ADS domains . *The connection between selection bias and other downstream issues (e.g., intervention effects, selective labels) underscores the importance of considering the full data generating process while diagnosing sources of bias impacting proxy labels.*

3.5.5 Confounding bias. In causal inference settings, confounding bias occurs when unmeasured variables influence both the treatment and outcome variable [75]. Confounding impacts ADS tasks when unobservables Z influenced past decisions D and downstream outcomes Y^* and Y (Figure 2.E) [75]. When confounding impacts ADS models , it is not possible to fully mitigate treatment effects and selective labels via traditional causal inference techniques [76]. Yet, confounding is *not* introduced by model unobservables Z in decision-independent settings because there is no arrow from decisions D to outcomes Y and Y^*. In these settings, unobservables may serve as an *opportunity for complementarity* between humans and models arising from asymmetric access to information [45, 47]. This differential role of unobservables across tasks with decision-dependent and decision-independent target variables underscores the importance of considering the structural causal model underlying the ADS task.

4 MODEL DEVELOPMENT

We now provide a framework for surfacing target variable modeling assumptions during predictive model development. We argue that predictive modeling for ADS involves two distinct steps: *measurement* and *prediction*. During the measurement step, tool developers construct a *measurement model* that operationalizes the target outcome of interest Y^* using readily available datasets. During the second step, tool designers train a *prediction model* that targets the *proxy outcome* returned by the measurement model. We now discuss each of these modeling steps in detail.

4.1 Measurement model

During the measurement step, the unobserved outcome of interest (Y^*) is approximated using historical data. This step involves establishing a *measurement hypothesis* ($\hat{Y}^*$) using observed information: covariates X, past decisions D, and one or more outcome proxies Y (Figure 1). In some settings, a subset of unobservables are available during model development, but unavailable in deployment. Such runtime confounders $Z_r \subseteq Z$ may be present. For example, when protected attributes (e.g., race, gender) are available during development, but not in deployment for legal reasons [19, 26]. Given information X, Z_r, D, Y recorded in existing data, we can construct a measurement model approximating Y^*:

$$\hat{Y}^* = F_m[X, Z_r, D, Y] \tag{1}$$

Unlike statistical models commonly used in machine learning contexts, a measurement model cannot be learned from past data because the target outcome Y^* is unobserved. Instead, F_m relies on *measurement assumptions* concerning the relationship between the unobserved outcome of interest and recorded information available for modeling. Therefore, it is not possible to assess the quality of $\hat{Y}^*$ by comparing against held-out data, as is common in prediction settings. Instead, evaluating measurement models requires a multifaceted approach, including assessments of construct validity, synthetic experiments, sensitivity analyses, and other evaluation strategies described in Section 4.3.

All predictive models in ADS introduce a measurement model. However, this model is often *implicitly defined* and makes tacit assumptions on the relationship between available data sources (X, Z_r, D, Y) and the target variable (Y^*). Table 1 provides a detailed list of the measurement models assumed by existing ADS approaches. This table reifies often-implicit measurement assumptions adopted by prior work. In the bottom three rows, we apply our taxonomy to workhorse methods used in machine learning [64, 69], quantitative social sciences [63], and bio-statistics [48] literature.

4.2 Prediction model

After establishing a measurement model to estimate Y^* given (X, Z_r, D, Y), tool designers then develop a *prediction model* for use in decision-support settings. This prediction model takes observed covariates (X) and predicts the measurement hypothesis ($\hat{Y}^*$) established during the preceding measurement step. Because Z_r and Y are unavailable during deployment, these are not included in the prediction model. Most often, prediction models do not assume human decisions D are available at runtime (i.e., algorithm-in-the-loop [38]). However, in some more nuanced decision-making workflows, models may also assume that human decisions are available at run-time as an additional input (i.e., [36, 62, 91, 96]). Given X and optionally D available at runtime, the prediction model estimates

Work	Measurement (F_m)	Prediction (F_p)	Assumptions	Bias
Gao et al. [36] Madras et al. [62] Wilder et al. [96] Tan et al. [91] Hilgard et al. [46]	$\hat{Y}^* = F_m[Y]$	$\hat{Y} = \hat{F}_p[X, D_H]$ Human decisions D_H available at runtime	Proxy and target variables are equivalent $Y^* = Y$	None
De-Arteaga et al. [24]	$\hat{Y}^* = F_m[X, D, Y]$, where $\hat{Y}^* = D$ *expert consistency* instances and Y otherwise	$\hat{Y} = \hat{F}_p[X]$ Human decisions D_H unavailable at runtime	Expert consistency assumption	Measurement error, Selection bias
Lakkaraju et al. [60]	$\hat{Y}^* = F_m[Y]$		Heterogeneous acceptance rates	Selection bias
Coston et al. [20]	$\hat{Y}^* = F_m[Y_d]$, where Y_d is a potential outcome		Causal identifiability conditions	Intervention effects
Coston et al. [18]	$\hat{Y}^* = F_m[Y_d, Z_r]$, where Y_d is a potential outcome		Causal identifiability conditions	Intervention effects, Unobservables
Wang et al. [94]	$\hat{Y}^* = F_m[Y]$, where Y error is group-dependent		Confident learning assumptions (see [70])	Measurement error
Label noise Menon et al. [64]	$\hat{Y}^* = F_m[Y]$, where Y class-conditional or positive and unlabeled	ERM with surrogate loss (see [69])	Weak separability	Measurement error
Latent Class Analysis McCutcheon [63]	$\hat{Y}^* = F_m[Y]$, where $Y = \{Y^1, ..., Y^K\}$ are independent factors	3-step LCA with covariates (see [93])	$Y^i \perp\!\!\!\perp Y^j \mid Y^*$	Measurement error
Hui-Walter Framework Hui and Walter [48]	$\hat{Y}^* = F_m[Y]$, where $Y = \{Y^1, ..., Y^K\}$ are diagnostic tests	N/A	Test Se/Sp identifiability assumptions	Measurement error

Table 1: Taxonomy of measurement and prediction approaches. Top: methods proposed in ADS literature. Bottom: methods applied in machine learning, social sciences, and bio-statistics. *Bias* specifies which sources of target variable bias the approach addresses.

the *measurement hypothesis* $\hat{Y}^*$:

$$\hat{Y} = \hat{F}_p[X, D] \tag{2}$$

Whereas a measurement model is *constructed via measurement assumptions*, the prediction model $\hat{F}_p$ is a *learned mapping* from X (and in some cases D) to the measurement hypothesis $\hat{Y}^*$. Therefore, it is appropriate to evaluate generalization of $\hat{F}_p$ to held-out data via the standard slate of evaluation metrics (e.g., accuracy, AU-ROC, or statistical fairness measures). Critically, this evaluation is conducted with respect to the measurement hypothesis established during the measurement step rather than the target outcome directly. Thus, showing strong performance of $\hat{F}_p$ *is not sufficient to claim a model generates valid predictions for the target outcome* Y^*.

4.3 Measurement model evaluation

Because the target outcome Y^* is unobserved, measurement model evaluation requires a holistic, multifaceted approach leveraging converging sources of evidence. Informed by methods used in statistics, quantitative social sciences, and learning sciences, we provide a recommended set of approaches for validating measurement models in ADS tasks.

4.3.1 Construct reliability and validity. Measurement theory offers a comprehensive set of criteria for assessing the quality of a measurement model. *Construct reliability* describes the degree to which a latent phenomena is consistently reflected by a measurement model (e.q. 1) over time. Threats to construct reliability have been well documented in settings in which target variables are assigned via subjective human annotations. In these settings, assignment of target outcomes can vary substantially based on rater identity [25, 25], context [74], and specification of the annotation protocol [80]. Construct validity describes the extent to which a measurement model adequately captures an unobserved phenomenon of interest. Thus, while construct reliability is roughly analogous to the notion of statistical variance in F_m, construct validity is analogous to statistical *bias* in F_m . We refer the reader to [49**?**] for a detailed discussion of sub-components of construct reliability and validity that pertain to risk assessment development and evaluation.

4.3.2 Outcome cross-validation. In many ADS settings, multiple proxies are available that are believed to be related to the target outcome of interest. In criminal justice settings, courts often track multiple recidivism-related outcomes (e.g., 2-year general and violent recidivism, failure to appear); in child welfare settings, government agencies may track substantiation of abuse allegations, acceptance for welfare services, agency re-referral, placement in foster care, and hospitalization [92]. When multiple reference outcomes are available, *outcome cross-validation* can be used to train a model to predict one proxy , then evaluate this model on a slate of *additional reference variables* that modelers expect may be reasonable proxies for the outcome of interest. If targeting a proxy *also* results in strong performance across other reference variables, this provides evidence suggesting that a proxy may serve as a suitable measurement model. Outcome cross-validation has been independently used by analyses of proxy outcomes in learning analytics [82], criminal justice [53], child welfare [24], and healthcare [71] domains. Special cases of outcome cross-validation map to sub-components of construct validity. A model demonstrates *predictive validity* if its predictions correlate with a reference outcome known to be related to the target phenomena of interest [42]. A model demonstrates *discriminant validity* if its predictions are *not* correlated with an unrelated reference outcome.

4.3.3 Sensitivity analyses. Sensitivity analyses enable assessing the degree of *measurement model misspecification* permissible before evaluation of a *prediction model* is invalidated. This analytic technique was traditionally developed for causal inference settings to estimate the magnitude of unobserved confounding (i.e., unobservables Z) necessary to invalidate a causal effect estimate [27, 84]. More recently, sensitivity analyses have been developed for predictive modeling settings. For instance, [31] proposed a sensitivity analyses framework that examines the degree of outcome measurement error permissible before fairness-related analyses are invalidated. Future work in ADS would benefit from sensitivity analysis frameworks that examine multiple sources of target variable bias in parallel.

4.3.4 Synthetic evaluation. A key limitation of leveraging real-world datasets for measurement model validation is that one never knows the actual relationship between Y and Y^* in naturalistic data. Model-level evaluations in ADS typically circumvent this issue via *synthetic evaluations* that test whether proposed approaches are robust to experimentally manipulated bias [20, 24, 64, 94]. Yet, one limitation of synthetic evaluations is that they require assuming a specific measurement error model. If the data generating process adopted by a synthetic evaluation does not reflect real-world conditions, this can lead evaluation blind-spots. This concern is salient because synthetic evaluations are often designed with bespoke data generating processes intended to highlight the specific challenge being addressed by the technique.

4.3.5 The Oracle Test. Chouldechova et al. [17] propose a conceptual tool called the *"Oracle Test"*, which can surface unforeseen sources of target variable bias. This thought experiment supposes that we have access to an oracle model that can predict a proxy Y with perfect accuracy. The key question posed by this test is: *"What concerns remain given access to an oracle?"*. Because we have a "perfect" prediction model (i.e., e.q. 2), remaining concerns are often related to measurement and validity. For example, Chouldechova et al. [17] surface concerns related to measurement error when they apply the Oracle Test to examine RAIs designed for ADS models deployed for child welfare. Green and Chen [39] also leverage the Oracle Test to argue that improvements to predictive accuracy do not equate to improved policy outcomes when competing factors in addition to risk (i.e., defendant liberty) play a role in judicial decision-making.

5 ASSESSING GAPS AND OPPORTUNITIES FOR EXPERIMENTAL RESEARCH

Our causal diagram can be used to assess the extent to which existing in vitro studies consider measurement error, intervention effects, and related challenges (Section 5.1). By mapping existing studies to corresponding sub-regions of our proposed diagram, we surface key blind-spots in our current understanding of human-AI decision-making. Our causal diagram can be used to evaluate whether experimental findings are likely to hold *ecological validity* in a given real-world ADS deployment (Section 5.2). More broadly, the methodology of causal diagrams can also serve as a scaffolding for structuring empirical knowledge beyond concerns related to target variable bias (Section 5.3).

5.1 Mapping existing experimental study designs to our causal diagram

To assess the extent to which existing studies examine factors related to target variable bias in their study design, we revisit a comprehensive literature review conducted by Lai et al. [57] through the lens of our causal diagram (Figure 1). Lai et al. [57] review over one hundred experimental studies of algorithm-assisted decision-making published in premiere venues between 2018 and 2021. Our follow-up analysis extends this review to studies published in 2022 at the same set of venues, in addition to other recently published pre-prints. We further limit selection criteria applied by Lai et al. [57] to studies examining human-AI decision-making in prediction-based decision-making settings (i.e., scope outlined in Section 3.1). Thus, we exclude studies included in the initial review with a focus on NLP-related tasks.

We find that 66 out of 72 (≈ 92%) studies satisfying our criteria conduct experimental evaluations focusing on a narrow sub-graph of our causal diagram. These studies investigate a modification to the joint human-AI decision-making process (i.e., the blue D_H and D_A region) using observed attributes X and an outcome proxy Y (Figure 3). Such studies assume that (1) the target variable and proxy are equivalent (i.e., no measurement error or validity concerns), (2) all predictive attributes of interest are observed by both the algorithm and the human (i.e., no model unobservables), and (3) decisions and outcomes are statistically independent (i.e., no intervention effects).

Six of the remaining studies we review examine different sub-regions of the causal diagram described in Figure 1. Table 2 groups these studies by the sub-region under study, including unobservables [45, 47], measurement error [37], selection bias [77], and omitted payoffs [32, 39]. While these studies offer early insight into how target variable bias can impact algorithm-assisted human

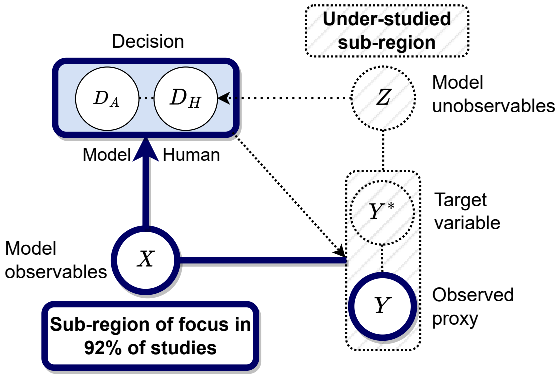

Figure 3: 66 of the 72 studies (≈ 92%) in our review examine a narrow sub-region of our proposed causal diagram.

Work	Setting	Sub-region of causal diagram
Hemmer et al. [45] Holstein et al. [47]	House price prediction	Unobservables: $D \leftarrow Z \rightarrow Y$
Gordon et al. [37]	Toxicity detection	Measurement error: $Y^* \rightarrow Y$
Peng et al. [77]	Hiring	Selection bias: $X \rightarrow D$
Green and Chen [39] Fogliato et al. [32]	Judicial	Omitted payoffs: $Q \rightarrow D_H$

Table 2: Experimental studies examining the under-studied sub-region provided in Figure 1

decision-making, our empirical understanding of these challenges remains limited compared to the joint human-AI decision region investigated by ≈ 92% of studies. Critically, no work in our review experimentally manipulated factors related to *intervention effects* or examined *multiple intersecting sources of bias* in parallel. Given the prevalence of compounding challenges in real-world settings, *this gap opens a broad space of open questions and future research opportunities* for research on human-AI decision-making.

5.2 Assessing ecological validity of in vitro studies

The gap we identify between in vivo challenges and in vitro studies (i.e., Figure 3) carries implications for the ecological validity of experimental studies. Threats to ecological validity may be most acute when findings from a controlled study conducted under simplified conditions are *generalized* to real-world ADS deployments in which multiple sources of target variable bias are present. In these settings, measurement error and intervention effects could impact whether findings gathered via controlled experiments also apply in more complex real-world conditions.

Fortunately, our causal diagram provides a tool for assessing whether findings from an *in vitro* study are likely to generalize to a given *in vivo* ADS tool deployment. The first step in this process involves mapping the decision support task to its corresponding DAG (e.g., identifying whether the task involves decision-dependent vs. decision-independent outcomes). Next, based on domain expertise, one can identify whether different sources of bias are likely to be relevant in the given real-world deployment. For instance, a model deployed to allocate tutoring resources (selective intervention) may need to account for measurement error in learning outcomes and intervention effects from historical tutoring decisions. In contrast, a model deployed for a perceptual assessment task (e.g., predicting current forest cover from limited image data) may not need to address these concerns. After identifying the DAG and relevant sources of bias, one can assess whether an experimental study is likely to generalize to this setting by examining whether the study used a similar prediction task (i.e., also tested decision-dependent or decision-independent outcomes).

To demonstrate how causal diagrams can be used to assess ecological validity of lab-based studies, consider a previous lab-based assessment conducted by Park et al. [73]. This study – which is sampled from the 66 studies covered by the blue sub-region of the causal diagram provided in Figure 3 – examines whether introducing a delay between when humans view observed features X and algorithmic recommendations D_H improves their performance on a perceptual jellybean counting task. Because the true quantity of jellybeans does not depend on the decision under consideration, this setting involves decision-independent outcomes. Further, the influence of human-only observed attributes Z and measurement error is limited in this task. Therefore, findings from this work may most readily generalize to real-world decision-making settings with limited interference from measurement error, model unobservables, and treatment effects.

5.3 Scaffolding a science of human-AI decision-making

Our work leverages causal diagrams to characterize sources of bias impacting target variables. However, beyond this focus, causal diagrams also offer a powerful scaffolding for studying other aspects of human-AI decision-making. While we model the target variable (Y^*) endogenously as a function of (X, Z, D, Y), one could also construct a causal diagram examining the joint human-AI decision-making process. For example, Green and Chen [39] specify such a DAG that models how humans weigh risk against other competing factors (e.g., culpability, value of defendant freedom) during pre-trial release decisions. The authors then experimentally verify a *hypothesized* edge in this DAG via a controlled online study. Through a series of such studies, it may be possible to develop a more generalized *theory* of AI-assisted human decision-making across decision-support tasks. This process of specifying, testing, and refining causal models is central to existing empirical disciplines, including psychology and sociology [75].

6 DISCUSSION

Our work surfaces a disconnect between the challenges that arise in *real-world deployments* of algorithmic systems versus *current research practices* adopted in the literature on human-AI decision-making (i.e., experimental study designs, modeling assumptions, measures of human-AI decision quality). Addressing this disconnect is critical in light of the grave real-world impacts introduced by target variable bias, including misallocation of medical resources [71] and perpetuation of historical bias in the criminal justice system [1, 5, 33], among many others (see Appendix A).

Our work provides a critical first step for addressing this disconnect by clarifying the relationship between measurement error, intervention effects, unobserved confounding, and selection bias via intuitive causal diagrams. Going forward, we hope that this framework will support more comprehensive assessment of modeling assumptions (Sections 4.1 and 4.2), rigorous evaluation of measurement-models (Section 4.3), and experimental human subjects studies that investigate the implications of bias in algorithm-assisted *human* decision-making (Section 5). However, our work is only the first step towards comprehensively addressing target variable bias in human-AI decision-making.

In particular, future research should develop holistic measures of decision-quality that reflect factors beyond statistical performance computed via a single outcome proxy. These measures should reflect both *process-oriented* considerations (i.e., how multiple decision-relevant factors are weighted [38], and adherence to procedural, interpersonal, and informational justice) in addition to *outcome-oriented* considerations (i.e., whether a decision led to a beneficial outcome). Where possible, outcome-related measures should draw upon *multiple decision-relevant proxies* to better account for limitations of adopting any single proxy in isolation. To date, human-AI decision-making research has primarily adopted outcome-oriented measures that hinge upon on a single potentially flawed proxy.

Our work also motivates exciting new lines of human-AI decision-making research. For instance, our review of experimental human subjects research (Section 5.1) identifies an open space of rich empirical questions to be investigated through future studies. We currently have an incomplete understanding of how measurement error impacts human trust and reliance on algorithmic predictions. Further, while counterfactual prediction methods have been proposed in the modeling literature, we currently have limited knowledge of how these methods can be leveraged to support more effective human decision-making practices. More generally, our causal framework provides a set of tools for (i) identifying open empirical questions, (ii) designing studies with robust ecological validity, and (iii) synthesizing findings from multiple experimental studies into a complete scientific understanding of human-AI decision-making. We hope that our work will raise awareness of target variable bias in the human-AI decision-making research community and spur efforts to better align research practices with the complex challenges encountered in real-world ADS deployments.

7 ACKNOWLEDGEMENTS

We thank Stevie Chancellor, Steven Dang, Maria De-Arteaga, Shamya Karumbaiah, Ken Koedinger, and attendees of the NeurIPS 2022 Workshop on Human-Centered Machine-Learning for their helpful feedback. This work was supported by an award from the UL Research Institutes through the Center for Advancing Safety of Machine Intelligence (CASMI) at Northwestern University, the Carnegie Mellon University Block Center for Technology and Society (Award No. 53680.1.5007718), and the National Science Foundation Graduate Research Fellowship Program (Award No. DGE-1745016). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the grantors.

REFERENCES

[1] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 838–849.

[2] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[5] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498* (2021).

[6] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*. PMLR, 62–76.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[8] Richard A Berk. 1983. An introduction to sample selection bias in sociological data. *American sociological review* (1983), 386–398.

[9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[10] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[11] Bradley Butcher, Chris Robinson, Miri Zilka, Riccardo Fogliato, Carolyn Ashurst, and Adrian Weller. 2022. Racial Disparities in the Enforcement of Marijuana Violations in the US. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 130–143.

[12] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[14] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016), 124–27.

[15] Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. 2022. Sample Efficient Learning of Predictors that Complement Humans. In *International Conference on Machine Learning*. PMLR, 2972–3005.

[16] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkat Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (forthcoming)*.

[17] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[18] Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. 2020. Counterfactual predictions under runtime confounding. *Advances in Neural Information Processing Systems* 33 (2020), 4150–4162.

[19] Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. 2020. Counterfactual predictions under runtime confounding. *Advances in Neural Information Processing Systems* 33 (2020), 4150–4162.

[20] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 582–593.

[21]]coston2022validity Amanda Lee Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. [n. d.]. SoK: A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.

[22] Robyn M Dawes, David Faust, and Paul E Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.

[23] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2018. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905* (2018).

[24] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648* (2021).

[25] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554* (2021).

[26] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2022. Multiaccurate proxies for downstream fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1207–1239.

[27] Iván Díaz and Mark J van der Laan. 2013. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics* 9, 2 (2013), 149–160.

[28] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 7, 4 (2016).

[29] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *arXiv preprint arXiv:2202.08821* (2022).

[30] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. PMLR, 160–171.

[31] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. 2020. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2325–2336.

[32] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[33] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 100–111.

[34] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[35] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.

[36] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614* (2021).

[37] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[38] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[39] Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[40] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12, 1 (2000), 19.

[41] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[42] David J Hand. 2004. Measurement theory and practice. *London: Arnold* (2004).

[43] Moritz Hardt and Michael P Kim. 2022. Backward baselines: Is your model predicting the past? *arXiv preprint arXiv:2206.11673* (2022).

[44] Harry H Harman. 1976. *Modern factor analysis*. University of Chicago press.

[45] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2022. On the Effect of Information Asymmetry in Human-AI Teams. *arXiv preprint arXiv:2205.01467* (2022).

[46] Sophie Hilgard, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, and David Parkes. 2021. Learning representations by humans, for humans. In *International Conference on Machine Learning*. PMLR, 4227–4238.

[47] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2022. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *arXiv preprint arXiv:2207.13834* (2022).

[48] Sui L Hui and Steven D Walter. 1980. Estimating the error rates of diagnostic tests. *Biometrics* (1980), 167–171.

[49] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

[50] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. 2020. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426* (2020).

[51] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*. PMLR, 2439–2448.

[52] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[53] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[54] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[55] Amanda Kube, Sanmay Das, and Patrick J Fowler. 2019. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 622–629.

[56] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 299–318.

[57] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[58] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[59] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[60] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 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*. 275–284.

[61] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[62] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems* 31 (2018).

[63] Allan L McCutcheon. 1987. *Latent class analysis*. Number 64. Sage.

[64] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *International conference on machine learning*. PMLR, 125–134.

[65] Smitha Milli, Luca Belli, and Moritz Hardt. 2021. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 714–722.

[66] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5323–5331.

[67] Sendhil Mullainathan and Ziad Obermeyer. 2017. Does machine learning automate moral hazard and error? *American Economic Review* 107, 5 (2017), 476–80.

[68] Sendhil Mullainathan and Ziad Obermeyer. 2019. *A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions*. National Bureau of Economic Research.

[69] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems* 26 (2013).

[70] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.

[71] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[72] Elizabeth L Ogburn and Tyler J Vanderweele. 2013. Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika* 100, 1 (2013), 241–248.

[73] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[74] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998* (2020).

[75] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[76] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

[77] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

[78] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

[79] Steve Pischke. 2007. Lecture notes on measurement error. *London School of Economics, London* (2007).

[80] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* 55, 2 (2021), 477–523.

[81] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[82] Napol Rachatasumrit and Kenneth R Koedinger. 2021. Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice. *International Educational Data Mining Society* (2021).

[83] Fred S Roberts. 1985. Measurement theory. (1985).

[84] Paul R Rosenbaum. 2005. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science* (2005).

[85] Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*. PMLR, 489–511.

[86] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[87] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[88] Vernon C Smith, Adam Lange, and Daniel R Huston. 2012. Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of asynchronous learning networks* 16, 3 (2012), 51–61.

[89] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, et al. 2022. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *arXiv preprint arXiv:2207.10192* (2022).

[90] Adith Swaminathan and Thorsten Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research* 16, 1 (2015), 1731–1755.

[91] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).

[92] Rhema Vaithianathan, Emily Kulick, Emily Putnam-Hornstein, and D Benavides-Prado. 2019. Allegheny family screening tool: Methodology, version 2. *Center for Social Data Analytics* (2019), 1–22.

[93] Jeroen K Vermunt. 2010. Latent class modeling with covariates: Two improved three-step approaches. *Political analysis* 18, 4 (2010), 450–469.

[94] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 526–536.

[95] Bridget E Weller, Natasha K Bowen, and Sarah J Faubert. 2020. Latent class analysis: a guide to best practice. *Journal of Black Psychology* 46, 4 (2020), 287–311.

[96] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

[97] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. 1–28.

8 APPENDIX

8.1 Descriptions of widely-studied outcome measurement error models

- **Uniform** error assumes that the target outcome is randomly corrupted by additive noise (i.e., $Y^* = Y + \epsilon$) [79]. This setting is called *classical measurement error* in statistics and economics and *symmetric label noise* in machine learning [2]. Because it is possible to learn an unbiased estimate for Y^* given proxy labels Y in uniform error settings [64], this error model poses fewer threats to validity than others discussed below.
- **Class-dependent** error assumes that positive and negative target outcomes are misclassified at different rates. As with uniform error„ measurement error in this setting is uncorrelated with covariates ($Y \perp\!\!\!\perp Y^*|X$) and model unobservables ($Y \perp\!\!\!\perp Y^*|Z$). This model is referred to as *asymmetric* or *class conditional label noise* in machine learning literature [85], and *nondifferential mismeasurement* in statistics and epidemiology [72]. In contrast to uniform error settings, training a model to predict a proxy (Y) impacted by class dependent error *will* lead to biased estimates for the target outcome (Y^*) when optimizing accuracy [64].
- **Feature-dependent** error occurs *differentially* across subpopulations based on covariates ($Y \not\perp\!\!\!\perp Y^*|X$) or model unobservables ($Y \not\perp\!\!\!\perp Y^*|Z$). This model is called *differential mismeasurement* in statistics and *feature-dependent label noise* in machine learning literature [34]. This setting is also called *group-dependent error* when the covariate in question is a protected attribute (e.g., gender, race) [94]. Group-dependent error has been tied to disparities in criminal justice [71] and medical [1] outcomes in real-world deployments of ADS tools.

Human-AI decision-making research also stands to benefit from existing *methodologies* designed to characterize measurement error in other disciplines. Latent Class Analysis (LCA) is an approach used in psychology and political science to identify latent subpopulations in data that are believed to carry an unobserved characteristic (e.g., personality, political ideology, or disease status) [95]. LCA estimates a set of conditional probabilities mapping multiple discrete *factors* (i.e., *proxies*) to a binary latent variable (e.g., *target outcome*). While LCA is tailored to discrete latent variables, other structural equation models (i.e., factor analysis [44]) are designed for continuous latent variables. Within biostatistics, the Hui-Walter framework is used to estimate the sensitivity and specificity of diagnostic tests in the absence of a gold standard [48]. Given multiple proxies, Hui-Walter can therefore be adapted to estimate the sensitivity and specificity of each proxy. Like all measurement models, LCA and Hui-Walter make assumptions on the relationship between the target outcome and its proxy. Table 1 states these assumptions in the context of our measurement model taxonomy.

Work	Domain	Bias Reported
Kleinberg et al. [53]	Judicial	Unobservables, selection bias, and outcome measurement error impacting pre-trial risk assessments
Fogliato et al. [33]		Measurement error introduced by adopting re-arrest as a proxy for re-offense
Bao et al. [5]		Selection bias and measurement error impacting by recidivism RAIs
Butcher et al. [11]		Measurement error in re-arrest proxy outcomes introduced by differential arrest rates among Black and white defendants
Kawakami et al. [52]	Child Welfare	Documents social worker concerns that measurement error and unobservables impact the quality of ADS predictions
Cheng et al. [16]		
Obermeyer et al. [71]	Medical	Measurement error arising from adopting *"cost of care"* as a health proxy
Mullainathan and Obermeyer [67]		Measurement error introduced when using medical records as a proxy for stroke outcomes
Mullainathan and Obermeyer [68]		Unobservables, selection bias, and measurement error in clinical decision support
Chalfin et al. [14]	Hiring	Omitted payoffs, measurement error, and selection bias arising in *teacher value-add* proxy used for educator hiring

Table 3: Documented examples of target variable bias impacting predictive models across numerous ADS domains.