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

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

Amanda Coston
acoston@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Zhiwei Steven Wu
zstevenwu@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Kenneth Holstein
kjholste@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, 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 commonly 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 for these nuanced phenomena 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.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **User studies**; • **Computing methodologies** → **Machine learning**.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594036

KEYWORDS

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

ACM Reference Format:
Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3593013.3594036

1 INTRODUCTION

A growing body of research aims to combine predictive machine learning (ML) models with human judgment to improve decision-making processes. In the machine learning community, researchers have proposed improvements to *ML models* to better address gaps in human judgment (e.g., [51, 65, 93, 100]). 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, 9–11, 15, 60, 64]). However, current evaluations of both model-level and human behavioral interventions typically assess the quality of human decisions, algorithmic predictions, and hybrid combinations of the two by comparing their accuracy on "ground-truth" labels that are readily available in existing data. This practice assumes that the labels targeted by predictive models serve as a reliable measure of the underlying goals and objectives of human decision-makers.

Yet in real-world deployments of algorithmic decision support (ADS) tools, labels are often imperfect proxies for the target outcomes of interest to human experts. While making decisions, content moderators frequently assess the "toxicity" of online comments [45] while physicians often consider the "cardiovascular disease risk" of patients [2]. "Toxicity" and "cardiovascular disease risk" are examples of latent constructs which are unobserved in data. Because observed labels (e.g., toxicity annotations and diagnostic test results) serve as indirect measurements of these phenomena [54], 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 [63], intervention effects [22], and selection bias [85], which interact with measurement error in nuanced ways depending on the nature of the specific 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* (TVB).[1] Following common terminology in statistics, we use the term "bias" to describe systematic differences between the target outcome of interest to human experts and its imperfect operationalization in available data. Thus, while TVB describes a broad conceptual difference between outcomes of interest and their observed proxies, this difference can be formally studied under existing statistical frameworks.

Target variable bias has been widely documented in real-world deployments of algorithmic systems [5, 12, 16, 18, 37, 56, 57, 70, 71, 74]. Predictive models impacted by target variable bias have contributed to unwarranted firing of teachers [16], perpetuated disparities in access to medical resources [74], and raised concerns among social workers investigating allegations of child abuse and neglect [18, 56]. Surprisingly, existing modeling efforts and human subjects experiments in the 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.**

Therefore, in this work, we bridge the divide 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 gaps in previously published modeling approaches and human subjects experiments, and (iii) providing guidelines for improved research practices going forward. In particular, we develop a *causal framework* which identifies the sources and implications of target variable bias in human-AI decision-making by examining the data generating process which gives rise to predictive model training datasets. Our framework enables us to distill ADS tasks studied in prior literature into their underlying structural components, and identify which sources of TVB (e.g., measurement error, intervention effects, selective labels) are of concern in a specific task. Using our framework, we identify opportunities to better address 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**. We use our framework to re-examine the design of prior human subjects experiments studying human-AI decision-making. Our analysis identifies systematic gaps in our current understanding of human-AI decision-making due to target variable bias.

[1]The term Target Variable Bias was introduced in [19, 35]. We use this as an umbrella term describing sources of statistical bias known to impact proxy labels in decision support tasks.

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 designed 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 [24, 44], methods have been proposed that learn to complement humans by adaptively routing decision instances [40, 65], leveraging heterogeneity in human and machine decision performance [17, 33, 93, 100], leveraging consistency in expert decisions [26], and adapting to [51] and training [69] 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 common 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 [10, 14, 39, 42, 61, 62]. This work investigates the potential for tools such as training protocols [13, 14, 61], explanations [11, 64], and other behavioral interventions [10, 39], to improve how humans make use of model outputs. While many online experimental studies have 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 [50, 52], measurement error [41], and omitted payoffs [43].

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 [57], selective labels [63], selection bias [25, 89], and intervention effects [22]. Additional work has examined the quality of proxy labels in decision support tasks. For example, Obermeyer et al. [74] 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 [16, 26, 57]. While this bias describes challenges specific to prediction (e.g., model unobservables, measurement error [57]), this term also applies when humans care about a broader set of decision-making factors beyond predictive risk [16, 42]. In this work, we use the lens of measurement and validity to examine systematic differences between target outcomes of interest to humans and proxy labels observed in data [54]. 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 [86].

2.3 Measurement and validity in algorithmic systems

Recent work has raised broader concerns regarding whether algorithmic systems successfully achieve their purported function [5, 23, 54]. Synthesizing concepts from measurement theory in the quantitative social sciences, Jacobs and Wallach [54] 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 identifies several biases in the outcome variable Y targeted by models, which we further characterize in this work. Coston et al. [23] highlight validity concerns impacting RAIs, including many discussed in § 2.2. Recent work has also surfaced validity issues in content moderation [41] and recommender systems [68, 92].

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 use causal diagrams to examine the relationship between measurement error and additional modeling challenges (i.e., § 2.2) that can impact the validity of prediction targets in real-world decision support settings. To our knowledge, our work offers the first holistic examination of how measurement error, unobservables, selection bias, intervention effects, and confounding interact to impact target variable validity in real-world ADS deployments.

3 FRAMEWORK

We now describe our framework scope (§ 3.1) and development process (§ 3.2) before introducing our causal diagram (§ 3.3). We then use our framework to map algorithmic decision support tasks to relevant sources of target variable bias (§ 3.4).

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 [74], criminal justice [31], child welfare [19], or real estate [52, 83] related outcomes); (ii) a subjective human annotation (e.g., perceived content toxicity [41]); or (iii) factual information (e.g., food nutrition [10]). In these settings, model predictions are combined with human decision-making, either by showing model predictions to a human (i.e., algorithm-in-the-loop [42]), who makes the final decision, or via a hybrid flow of agency (e.g., deferral-based learning [65], learning with bandit feedback [40]). 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 [78], are tools specifically designed for this purpose. 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 across settings described by the diagram [78]. Our framework introduces 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 [13, 14, 61], decision-making process [43], and workflow [42] of the human decision-makers using the predictive model. While prior work has examined these factors in detail [13, 14, 42, 43, 61], our framework foregrounds factors most salient for understanding target variable bias. In § 5.3, we outline how our approach can be extended to systematically examine a broader set of components beyond target variables in human-AI decision-making 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 Table 3 in 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 the relationship between predictors, decisions, target variables, and their proxies in ADS tasks.

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 [71]) 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 [56]. 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) [52]. 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, post-deployment decisions result from humans incorporating predictions into their decision-making (i.e., algorithm-in-the-loop [42]). In other cases, the joint

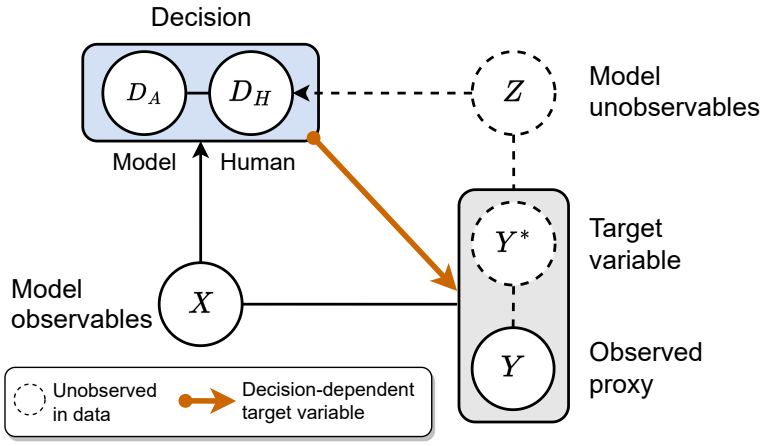

Figure 1: Our causal diagram represents a space of causal graphs, spanning different possible relationships between predictors, decisions, target variables, and their proxies in algorithmic decision support tasks. Edges with directionality that can vary across ADS settings are indicated via undirected 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.

decisions result from a learned combination of D_H and D_A [17, 33, 40, 51, 65, 93, 100].

Target variables. The node Y^* describes the unobserved target variable of interest to human decision-makers. For example, a model might be introduced to weigh the 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" [74], "re-arrest" [35], "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 § 3.4.1).

Edges. We now describe the space of possible relationships connecting nodes in ADS tasks. Covariates and model 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^*, Y). For example, enrollment in a medical treatment program can increase medical costs (Y) while also improving patient health (Y^*). We show this relationship via the directed arrow $D \rightarrow Y, Y^*$.

The direction of causality between covariates (X), unobservables (Z), and prediction targets (Y and Y^*) can vary across ADS domains. In Figure 1, we convey this ambiguity via undirected edges. Causal diagrams for prediction tasks often show covariates (X) and unobservables (Z) contributing to downstream outcomes (Y, Y^*) via a domain-specific causal pathway [7]. In our diagram, this flow of information would be communicated via directed edges from X to (Y, Y^*), and from Z to (Y, Y^*). However, in some cases, the causal pathway can be *reversed* [47]. For example, this is possible if a patient's unobserved disease status (Y^*) contributes to their medical history (X) or real-time test results (Z). Therefore, bidirectional edges shown in Figure 1 map to directed edges with directionality that varies depending on the domain. [2]

3.4 Mapping algorithmic decision support tasks to sources of target variable bias

We now leverage our causal framework to identify sources of target variable bias that can impact predictive models in algorithmic decision support tasks. We begin by introducing two distinct *regimes of ADS tasks* described by our generalized diagram shown in Figure 1; those with: (1) decision-dependent target variables, and (2) decision-independent target variables. **ADS tasks with decision-dependent target variables are subject to more sources of TVB than those with decision-independent target variables**. While it is possible to define many other specific regimes of our generalized diagram (e.g., different directions of causality between (Y, Y^*) and X or Z, or different flows of agency between D_A and D_H), we introduce the distinction between decision-dependent and independent target variables here because it is useful for identifying task-specific sources of TVB.

ADS tasks with **decision-dependent target variables** occur when *the decision informed by an algorithm also impacts the downstream outcomes* Y and Y^*. Real-world ADS deployments often involve prediction tasks with decision-dependent target variables. For example, re-arrest is only observed among defendants released on bail [57], while child welfare screening decisions can influence

[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 [34].

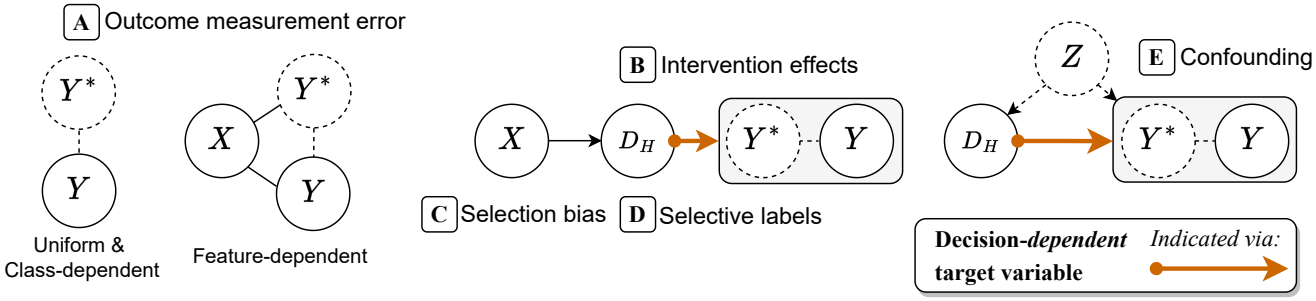

Figure 2: Sub-graphs of the diagram in Figure 1 introducing statistical biases that impact the target variable Y^*. *Outcome measurement error* (A) can occur in settings with both decision-dependent and independent target variables. In decision-dependent settings, *intervention effects* (B), *selection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

the risk of adverse care outcomes [22]. More generally, 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 target-variables include the orange arrow from D to Y and Y^* shown in Figure 1. [3]

In contrast, the target variable is *not* influenced by the proposed decision in ADS tasks with **decision-independent target variables**. ADS are frequently deployed in the real world with the goal of informing decisions that can change the predicted outcomes. However, lab-based experimental studies of human-AI decision-making often conduct evaluations via ADS tasks with decision-independent target variables. For instance, studies have examined models that predict factual content (e.g., food nutrition [10]) and perceptual information (e.g., counts of objects [76], geometric shapes [101]). These tasks are decision-independent because the prediction target (i.e., food nutrition, geometric shape) is *not* influenced by the prediction made by a human and/or model. [4] ADS tasks in the decision-independent regime do not contain the arrow from D to Y and Y^* in Figure 1.

We now introduce five sources of target variable bias relevant in ADS tasks. **Outcome measurement error is of concern in both decision-dependent *and* decision-independent regimes, while intervention effects, selective labels, selection bias, and confounding bias are only relevant in decision-dependent regimes.**

3.4.1 Outcome measurement error. Human experts 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* [46, 54]. 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, 37], 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 policymakers (Y^*) and its operationalization by a proxy (Y). This challenge has been extensively documented in judicial [12, 35], child welfare [18, 56], and hiring [16] ADS domains.

Because proxy labels impacted by measurement error offer an incomplete reflection of the actual goals of human decision-makers, they serve as an incomplete measure of human-AI decision quality. Therefore, before adopting a proxy as a measure of human-AI decision quality, it is critical to assess whether it serves as a satisfactory approximation of the target variable of interest to humans. Measurement theory in the quantitative social sciences provides tools to conduct this assessment by weighing the *construct validity* and *reliability* of observed labels [46, 54] (see § 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 A.1). Outcome measurement error is of concern in decision-dependent and independent regimes because observed labels can be subject to construct validity and reliability concerns in both settings.

3.4.2 Intervention effects. In many ADS tasks, decisions serve as *risk mitigating interventions* intended to improve the chances of a favorable policy-relevant outcome [6, 22, 59]. 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, 22, 59]. This practice can be traced back to formulation of ADS as a prediction-policy problem [58], 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

[3] This regime maps directly to the *"predictive optimization"* setting recently studied by Wang et al. [96] and the discussion of predictive model validity provided by Coston et al. [23].

[4] ADS tasks in which labels are assigned via human annotations also fall within the decision-independent regime. In these tasks, the ratings of human annotators (Y) serve as a proxy for the broader construct of interest (Y^*) of interest in the model deployment setting (e.g., comment "toxicity" or "hate speech" [41]).

the Allegheny Family Screening Tool (AFST) introduced the tool with the goal of assessing *"latent risk"* of maltreatment prior to county child welfare interventions [94]. Similarly, predictive models have been introduced in educational settings to identify students at-risk of failing given no tutoring resources [90]. 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. [22] demonstrate that predicting observed labels in SI settings systematically underestimates the risk for high-risk individuals who would respond most favorably to the intervention. This source of bias is only relevant in the decision-dependent regime because intervention effects are introduced by the connection $D \rightarrow Y, Y^*$.

3.4.3 Selective labels. Another challenge introduced by the connection $D \rightarrow Y, Y^*$ in the decision-dependent regime is *selective labels* (Figure 2.D). This bias has been widely discussed in connection to pre-trial risk assessments, where recidivism-related proxy outcomes (e.g., re-arrest, failure to appear) are only observed among defendants released on bail [5, 37, 57, 63]. 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 [25]. 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 data generating process.

3.4.4 Selection bias. This bias, which occurs when covariates (X) or model 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 [55]. While selection bias can cause challenges in any setting in which data is collected non-randomly [49], this challenge is compounded in decision-dependent outcome tasks because the connection $X \rightarrow D_H \rightarrow Y^*, Y$ causes selection effects to cascade to selective observation of outcomes Y and Y^*. 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.

3.4.5 Confounding bias. In causal inference settings, this bias occurs when unmeasured variables influence both the treatment and response variable [78]. Confounding impacts *ADS tasks* when unobservables influenced past decisions and downstream outcomes (Figure 2.E) [78]. When confounding impacts ADS models, it is not possible to fully mitigate treatment effects and selective labels via traditional causal inference techniques [79]. Yet, confounding is *not* introduced by model unobservables Z in decision-independent tasks because there is no arrow from D to Y and Y^*. In these tasks, unobservables may serve as an *opportunity for complementarity* between humans and models arising from asymmetric access to information [50, 52]. Therefore, by mapping an ADS task to its underlying causal diagram and identifying the appropriate task regime, it is possible to identify whether model unobservables pose a treat or opportunity for a given ADS deployment.

4 MODEL DEVELOPMENT

We now provide a framework for specifying target variable 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 variable 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 from the causal diagram in Figure 1. This step involves establishing a *measurement hypothesis* ($\hat{Y}^*$) using observed information: covariates X, past decisions D, and one or more outcome proxies Y. In some settings, a subset of unobservables are available during model development, but unavailable in during deployment. Such runtime confounders $Z_r \subseteq Z$ can occur when protected attributes (e.g., race, gender) are available during development, but not during deployment for legal purposes [21, 29]. Given information X, Z_r, D, Y recorded in existing data, we can construct a measurement model approximating the target variable 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,

Work	Measurement (F_m)	Prediction (F_p)	Assumptions	Bias Mitigated
Gao et al. [40] Madras et al. [65] Wilder et al. [100] Tan et al. [93] Hilgard et al. [51]	$\hat{Y}^* = F_m[Y]$	$\hat{Y} = \hat{F}_p[X, D_H]$ Human decisions D_H available at runtime	Proxy and target variables are equivalent $Y^* = Y$	None
De-Arteaga et al. [26]	$\hat{Y}^* = F_m[X, D, Y]$, 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. [63]	$\hat{Y}^* = F_m[Y]$		Heterogeneous acceptance rates	Selection bias
Coston et al. [22]	$\hat{Y}^* = F_m[Y_d]$, where Y_d is a potential outcome		Causal identifiability conditions	Intervention effects
Coston et al. [20]	$\hat{Y}^* = F_m[Y_d, Z_r]$, where Y_d is a potential outcome		Causal identifiability conditions	Intervention effects, Confounding
Wang et al. [97]	$\hat{Y}^* = F_m[Y]$, where Y error is group-dependent		Confident learning assumptions (see [73])	Measurement error
Label noise Menon et al. [67]	$\hat{Y}^* = F_m[Y]$, where Y class-conditional or positive and unlabeled	ERM with surrogate loss (see [72])	Weak separability	Measurement error
Latent Class Analysis McCutcheon [66]	$\hat{Y}^* = F_m[Y]$, where $Y = \{Y^1, ..., Y^K\}$ are independent factors	3-step LCA with covariates (see [95])	$Y^i \perp\!\!\!\perp Y^j \mid Y^*$	Measurement error
Hui-Walter Framework Hui and Walter [53]	$\hat{Y}^* = F_m[Y]$, 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.

synthetic experiments, sensitivity analyses, and other evaluation strategies described in § 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 [67, 72], quantitative social sciences [66], and bio-statistics [53] literature. The *Bias Mitigated* column of Table 1 refers to the source of TVB addressed by the modeling technique. For instance, we mark "None" for Wilder et al. [100] and Madras et al. [65] because these approaches are not designed to mitigate any TVB sources listed in § 3.4.

4.2 Prediction model

After establishing a measurement model to estimate Y^* given (X, Z_r, D, Y), tool designers then train 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 [42]). 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., [40, 65, 93, 100]). Given X and optionally D available at runtime, the prediction model estimates 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 ($\hat{Y}^*$) rather than the target outcome (Y^*) 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

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 [28, 28], context [77], and specification of the annotation protocol [82]. 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 [54]. We refer the reader to [23] 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 domains, multiple proxies are available that are believed to be related to the target outcome of interest. In the criminal justice domain, courts often track multiple recidivism-related outcomes (e.g., 2-year general and violent recidivism, failure to appear). In the child welfare domain, government agencies may track substantiation of abuse allegations, acceptance for welfare services, agency re-referral, placement in foster care, and hospitalization [94]. 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 domain experts 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 [84], criminal justice [57], child welfare [26], and healthcare [74]. Special cases of outcome cross-validation map to sub-components of construct validity. For example, a model demonstrates *predictive validity* if its predictions correlate with a reference outcome known to be related to the construct of interest [46]. A model demonstrates *discriminant validity* if its predictions are *not* correlated with a conceptually distinct 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 technique has traditionally been applied in causal inference settings to estimate the magnitude of unobserved confounding necessary to invalidate a treatment effect estimate [30, 87]. More recently, sensitivity analyses have been developed for predictive model evaluation. For instance, Fogliato et al. [35] 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 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* which test whether proposed approaches are robust to experimentally manipulated bias [22, 26, 67, 97]. Yet, synthetic evaluations 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 to overconfidence in model performance in more realisitic settings. 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. [19] 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 with perfect accuracy. The key question posed by this test is: *"What concerns remain given access to such an oracle?"* Because we have a "perfect" prediction model (e.q. 2), remaining concerns are often related to measurement and validity (e.q. 1). For example, Chouldechova et al. [19] surface concerns related to measurement error when they apply the Oracle Test to examine RAIs designed for ADS models deployed in the child welfare domain. Green and Chen [43] also leverage the Oracle Test by arguing that improvements to predictive accuracy do not equate to improved public policy outcomes when competing factors in addition to risk (i.e., defendant liberty) are overlooked.

5 ASSESSING GAPS AND OPPORTUNITIES FOR EXPERIMENTAL RESEARCH

In this section, we leverage our framework to assess the extent to which existing lab-based studies consider sources of target variable bias (§ 5.1). Our analysis finds systematic gaps in our current understanding of human-AI decision-making in light of TVB. We then show how our framework can be used by researchers to assess threats to the ecological validity and generalizability of lab-based studies (§ 5.2). We conclude by discussing opportunities to use our methodology to explore a broader space of open challenges in human-AI decision-making research (§ 5.3). In Appendix A.2, we provide a resource that helps researchers apply our causal framework to examine the design and ecological validity of several experimental human-AI decision-making studies.

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. [60] through the lens of our causal diagram (Figure 1). Lai et al. [60] review over

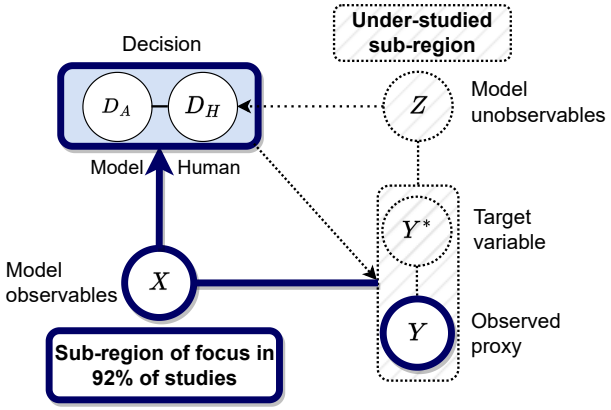

Figure 3: 66 of the 72 studies (≈ 92%) in our review examine a narrow sub-region of our proposed causal diagram.

Work	Setting	Sub-region of causal diagram
Hemmer et al. [50] Holstein et al. [52]	House price prediction	Unobservables: $D \leftarrow Z \rightarrow Y$
Gordon et al. [41]	Toxicity detection	Measurement error: $Y^* \rightarrow Y$
Peng et al. [80]	Hiring	Selection bias: $X \rightarrow D$
Green and Chen [43] Fogliato et al. [36]	Judicial	Omitted payoffs: $Q \rightarrow D_H$

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

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 recent pre-prints. We further limit selection criteria applied by Lai et al. [60] to studies examining prediction-based decision-making settings (i.e., scope outlined in § 3.1). Thus, we exclude studies included in the initial review with a focus on NLP-related tasks.

Our analysis finds 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 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), (2) all predictors are observed by both the algorithm and the human (i.e., no model unobservables), and (3) decisions and outcomes are unrelated (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 of focus, including unobservables [50, 52], measurement error [41], selection bias [80], and omitted payoffs [36, 43]. While these studies offer early insight into how target variable bias can impact algorithm-assisted human 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 opportunities for human-AI decision-making research**.

5.2 Assessing the ecological validity of lab-based studies

The gap we identify between real-world challenges and lab-based 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 a lab-based study are likely to generalize to a given real-world ADS tool deployment. The first step in this process involves mapping the ADS task to its corresponding regime identified in § 3.4 . 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 (i.e., a decision-dependent task) may need to account for mismeasured learning outcomes and intervention effects from prior tutoring program enrollment. In contrast, a model deployed for a perceptual assessment task (e.g., predicting current forest cover from satellite imagery [98]; a task with decision-independent outcomes) may not need to address these concerns. After identifying the appropriate ADS regime 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 (e.g., 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. [76]. 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_A improves their performance on a perceptual jellybean counting task. Because the true quantity of jellybeans does not depend on the decision under consideration, this study involves a task from the decision-independent outcome regime. 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 outcome measurement error, model unobservables, and intervention 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, such as the joint human-AI decision-making process (i.e., the D node in our framework). For example, Green and Chen [43] specify a causal diagram that models how judges 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 causal diagram 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 [78].

6 DISCUSSION

Our work surfaces a disconnect between the challenges that arise in *real-world deployments* of algorithmic systems versus *current research practices* (i.e., experimental study designs, modeling assumptions, measures of human-AI decision quality) adopted in the human-AI decision-making literature. Left unaddressed, current gaps in this literature can amount to substantive downstream harms. For instance, while prior studies of real-world ADS tool deployments have surfaced patterns of *apparent* human under-reliance arising from imperfect prediction targets [18, 56, 91], no experimental human subjects studies to date have examined how to disentangle warranted skepticism in a misaligned model versus *unwarranted* under-reliance due to algorithm aversion. Absent such knowledge, organizations may continue to pressure domain experts to rely upon flawed predictive models [56], which have been shown to misallocate of medical resources [74] and perpetuate historical patterns of bias [1, 5, 37] (see Table 3 in Appendix A for additional examples of real-world harms introduced by TVB).

Our work provides a critical first step for addressing this disconnect by clarifying the relationship between measurement error, intervention effects, unobserved confounding, selective labels, and selection bias via intuitive causal diagrams. Going forward, we hope that this framework will support more comprehensive assessment of modeling techniques (§ 4) and empirical human subjects studies (§ 5) designed to facilitate human-AI decision-making. However, further work is needed to gain a comprehensive understanding of the sources and implications of target variable bias in human-AI decision-making research.

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 [42], 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. While this practice is standard in disciplines such as learning sciences, diagnostic medical testing, and psychology, 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 prior modeling approaches finds that, while many techniques have been designed to address a subset of model reliability challenges (Table 1), few examine how various sources of target variable bias compound in real-world deployment scenarios. Additionally, our review of experimental human subjects research provides a set of tools for (i) identifying open empirical questions (i.e., Figure 3), (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.

ACKNOWLEDGMENTS

We thank Stevie Chancellor, Steven Dang, Maria De-Arteaga, Shamya Karumbaiah, Ken Koedinger, and annonymous reviewers for their helpful feedback. We acknowledge support 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). ZSW is supported in part by the NSF FAI (Award No. 1939606), a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, an Okawa Foundation Research Grant, and a Mozilla Research Grant.

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] Keaven M Anderson, Patricia M Odell, Peter WF Wilson, and William B Kannel. 1991. Cardiovascular disease risk profiles. *American heart journal* 121, 1 (1991), 293–298.

[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, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 95–106.

[16] 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.

[17] 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.

[18] 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)*.

[19] 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.

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

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

[22] 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.

[23] Amanda Lee Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. SoK: A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. *First IEEE Conference on Secure and Trustworthy Machine Learning* (2023).

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

[25] 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).

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

[27] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[28] 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).

[29] 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.

[30] 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.

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

[32] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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).

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

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

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

[49] James J Heckman. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* (1979), 153–161.

[50] 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).

[51] 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.

[52] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–20.

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

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

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

[56] 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.
[57] 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.
[58] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.
[59] 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.
[60] 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).
[61] 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.
[62] 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.
[63] 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.
[64] 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.
[65] 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).
[66] Allan L McCutcheon. 1987. *Latent class analysis*. Number 64. Sage.
[67] 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.
[68] 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.
[69] 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.
[70] Sendhil Mullainathan and Ziad Obermeyer. 2017. Does machine learning automate moral hazard and error? *American Economic Review* 107, 5 (2017), 476–80.
[71] 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.
[72] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems* 26 (2013).
[73] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.
[74] 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.
[75] Elizabeth L Ogburn and Tyler J Vanderweele. 2013. Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika* 100, 1 (2013), 241–248.
[76] 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.
[77] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998* (2020).
[78] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
[79] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[80] 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.
[81] Steve Pischke. 2007. Lecture notes on measurement error. *London School of Economics, London* (2007).
[82] 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.
[83] 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.
[84] 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).
[85] Ashesh Rambachan and Jonathan Roth. 2019. Bias in, bias out? Evaluating the folk wisdom. *arXiv preprint arXiv:1909.08518* (2019).
[86] Fred S Roberts. 1985. Measurement theory. (1985).
[87] Paul R Rosenbaum. 2005. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science* (2005).
[88] 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.
[89] 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.
[90] 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.
[91] Megan T Stevenson and Jennifer L Doleac. 2022. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440* (2022).
[92] 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).
[93] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).
[94] 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.
[95] Jeroen K Vermunt. 2010. Latent class modeling with covariates: Two improved three-step approaches. *Political analysis* 18, 4 (2010), 450–469.
[96] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2022. Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *Available at SSRN* (2022).
[97] 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.
[98] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
[99] 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.
[100] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).
[101] 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.
[102] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A APPENDIX

A.1 Descriptions of widely-studied outcome measurement error models

- **Uniform** error assumes that the target outcome is randomly corrupted by additive noise (i.e., $Y^* = Y + \epsilon$) [81]. This setting is also sometimes called *classical measurement error* in statistics and economics. Because it is possible to learn an unbiased estimate for Y^* given proxy labels Y in uniform error settings [67], this error model poses fewer threats to validity than others discussed below.

Work	Domain	Bias Reported
Kleinberg et al. [57]	Judicial	Unobservables, selection bias, and outcome measurement error impacting pre-trial risk assessments
Bao et al. [5]		Selection bias and measurement error impacting by recidivism RAIs
Butcher et al. [12]		Measurement error in re-arrest proxy outcomes introduced by differential arrest rates among Black and white defendants
Kawakami et al. [56] Cheng et al. [18]	Child Welfare	Documents social worker concerns that measurement error and unobservables impact the quality of ADS predictions
Obermeyer et al. [74]	Medical	Measurement error arising from adopting *"cost of care"* as a health proxy
Mullainathan and Obermeyer [70]		Measurement error introduced when using medical records as a proxy for stroke outcomes
Mullainathan and Obermeyer [71]		Unobservables, selection bias, and measurement error in clinical decision support
Chalfin et al. [16]	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.

- **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 co-variates ($Y \perp\!\!\!\perp Y^* | X$) and model unobservables ($Y \perp\!\!\!\perp Y^* | Z$). This model is referred to as *asymmetric* or *class conditional label noise* in machine learning literature [88], and *nondifferential mismeasurement* in statistics and epidemiology [75]. 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 [67].
- **Feature-dependent** error occurs *differentially* across subpopulations based on co-variates ($Y \not\perp\!\!\!\perp Y^* | X$) 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 [38]. This setting is also called *group-dependent error* when the covariate in question is a protected attribute (e.g., gender, race) [97]. Group-dependent error inherits modeling challenges arising in the class dependent case, and has been tied to disparities in criminal justice [74] 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) [99]. 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 [48]) 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 [53]. 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.

A.2 Extended review of prior experimental studies through the lens of our causal framework

In this section, we provide a resource to help researchers examine factors related to target variable bias during the design and evaluation of experimental human-AI decision-making studies. We provide a detailed examination of several studies included in our review [10, 32, 41, 42, 52, 102]. For each study, we identify (1) the **sub-region of focus**, and (2) the **ADS regime** used in the experimental evaluation.

- The **sub-region of focus** describes the primary nodes and edges considered in the experimental design and evaluation of the work (e.g., regions shown in Figure 3). This region can be determined by the description of the experimental design (i.e., conditions and RQs), task, and methods provided by the authors. For example, works often report the co-variates used to train a model (X), proxy label (Y), and experimental manipulation of focus in the study. For many studies included in our review, the experimental manipulation involves a modification to the joint decision region of our diagram (D) in the form of explanations [102], cognitive forcing functions [10], model accuracy [32], or other behavioral interventions.
- The **ADS regime** describes the data generating process that gave rise to the dataset used to train the predictive model examined in the experimental evaluation. Our causal framework contains two specific ADS regimes: those with (1) decision-independent target variables and (2) decision-dependent target variables. In contrast to the sub-region of focus, the ADS regime is *implicit* in the description of prior studies. This is because the majority of prior studies do not

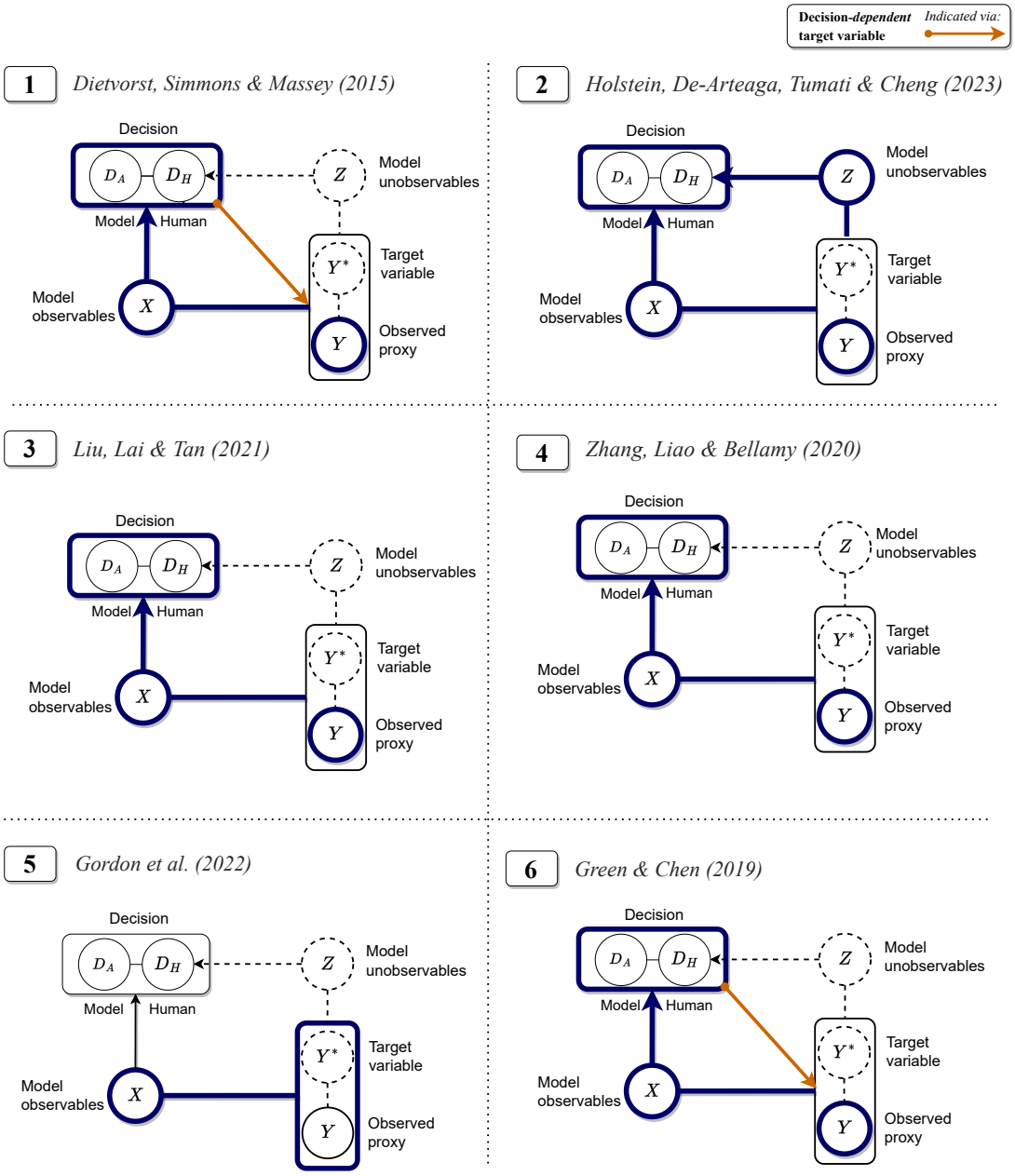

Figure 4: Causal diagrams for six of the studies included in our review of prior human-AI decision-making experiments discussed in § A.2. Sub-regions of focus in the study are shown with bold blue borders. The arrow connecting the D node with target variables Y and Y^* is omitted in tasks falling under the decision-independent regime.

explicitly discuss factors related to outcome measurement error, unobservables, or treatment effects that may be relevant in the task design.

A.2.1 Study 1: Dietvorst et al. [32]. In this study, Dietvorst et al. [32] popularize the term *algorithm aversion* by finding that "participants more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake." This study instructed participants to play the part of an MBA admissions officer by predicting the percentile the student would rank among their peers given application information such as undergraduate degree, GMAT scores, interview quality, essay quality, work experience, average salary, and parents' education. The primary experimental manipulation studied whether participants would elect to use

human judgement versus a statistical model given different information about their relative performance. We show the sub-region of focus and ADS regime in Figure 4.1.

- **Sub-region of focus:** D, X, Y. This study focuses on the sub-region with joint decisions D, co-variates X and outcome proxies Y. We include co-variates (X) because the authors list an explicit set of features that are provided to both the human and the model. We include the joint human-model decision region (D) because the experimental treatment alters participant awareness of human and model performance differences. We include proxy labels (Y) because the authors describe an outcome variable of student "success", defined as an average of multiple performance measures (GPA, respect of fellow students, and prestige of employer upon graduation). We do not include Y^* because the authors do not examine additional operationalizations of "success" or "student performance" that could be possible in this admissions setting. We do not include unobservables Z because the authors do not examine other factors (e.g., student demeanor, personal connections) that might be available to an admissions officer but not a model. We do not include the edge connecting decisions D and outcomes Y, Y^* because the authors do not examine the impact of predictions and admissions decisions on downstream student performance.
- **ADS Regime: decision-dependent target variable.** In this setting, the decisions of admissions offers determine which students are admitted to the graduate program, and, consequentially, which students have academic performance outcomes available. Recall from § 3.4 that this is a *selective opportunity* setting because we only observe outcomes for students provided the enrollment opportunity. As a result, confounding and selection bias are relevant in this modeling task, in addition to outcome measurement error. In real-world deployments of predictive models for admissions decisions, unobservables and outcome measurement error may impact the ADS deployment due to private information available to a loan officer and alternate definitions of "academic success" or "academic performance" that may be relevant in this setting. **As a result, findings from this study may be most likely to generalize to other decision-dependent target variable tasks (e.g., financial loan approvals, commercial job hiring decisions, or pre-trial release decisions).**

A.2.2 Study 2: Holstein et al. [52]. In this study,the authors examine model unobservables as a potential source of complementary in an AI-assisted house price prediction task. Participants were shown a set of "Facts and Features" about homes (e.g., year built, type of heating, number of bathrooms, zoning classification) and asked to predict the house's sale price. These facts corresponded to tabular features available in the training data. Three of the eight features were removed during model training to introduce synthetic unobservables, and experimental conditions varied how participants were prompted to consider these unobservables during their decision-making. We show the sub-region of focus and ADS regime in Figure 4.2.

- **Sub-region of focus:** Z, D, X, Y. This study focuses on the sub-region with joint decisions D, model observables X, **model unobservables** Z, and outcome proxies Y. We include model observables (X) because the authors list an explicit set of features that were provided to both the human and the model. We include the joint human-model decision region (D) because the experimental treatment involved different participant prompts for considering unobservables during their decisions. We include the unobservables region (Z) because the authors explicitly omit predictive features from the model during training, but provide these to participants at decision time. We include the proxy label (Y) because the authors list a predictive outcome of house sale price. However, we do not include Y^* because the authors do not examine other potential operationalizations of "house worth" possible in this task (e.g., the amount *a participant would pay* for a house versus its actual market sale price). We do not include the edge connecting decisions D and outcomes Y, Y^* because the authors do not examine the impact of price predictions on downstream sales.
- **ADS Regime: decision-independent target variable.** In this setting, the sale price predictions of participants does not impact downstream house sale prices. Therefore, we list this task as decision-independent target variable. While it is conceivable that loan officer, real estate agent, or online platform price predictions could impact house sale prices (e.g., Zestimates) in similar settings, this is **not** the case in this particular evaluation because there is not a decision being informed by the model that directly impacts observed prices. In particular, the historical data available for model training lists a full set of houses and their corresponding prices, with no prior human decisions/price predictions that might have impacted the price. Because observed prices are not connected to the prediction task in this study, we list this as decision-independent. Therefore, while outcome measurement error could be a concern in this setting due to differing notions of "house quality", selection bias, confounding, selective labels, and treatment effects are not a concern in this evaluation. **As a result, findings from this study may be most likely to generalize to other decision-independent target variable tasks (e.g., nutrient content prediction, forest cover prediction) and may be less likely to generalize to real-world predictive model deployments with decision-dependent outcomes.**

A.2.3 Study 3: Liu et al. [64]. This study examines whether interactive explanations and out-of-distribution examples can foster human-AI complementary. Out-of-distribution examples refers to a setting in which the human-AI team makes decisions involving instances from a distribution that differs in composition from the model training dataset. The authors experimentally manipulate (1) sources of distribution shift and (2) presentation of interactive explanations. The authors conduct evaluations via recidivism prediction tasks (see Study 6 below) and an occupation classification task in which participants predict an individual's occupation given a written biography drawn from the BIOS dataset. We show the

sub-region of focus and ADS regime for the occupation prediction task in Figure 4.3.

- **Sub-region of focus:** D, X, Y. This study focuses on the sub-region with joint decisions (D), model observables (X), and outcome proxies (Y). We include model observables (X) because participants were shown a written biography about each person drawn from the BIOS dataset [27]. We include decisions D because the authors experimentally manipulate the explanation type and data distribution and examine impacts on human-AI decision quality. We include the proxy label (Y) because the authors list a prediction target involving the *reported occupation* of an individual in the dataset (e.g., psychologist, physician, surgeon, teacher, and professor). We do not include Y^* because the *reported occupation* of individuals in the BIOS data can overlook reporting bias or multiple professions (e.g., physician and professor), which is not examined in the experimental manipulation. We do not include Z because the study participants and model were both given access to the same biography information. We do not include the edge connecting decisions D and outcomes Y, Y^* because participant guesses do not influence the occupation of individuals in the BIOS data.
- **ADS Regime: decision-independent target variable.** Because participant responses do not influence the occupations of individuals in the dataset, this is a task with decision-independent target variables. The authors do examine selection bias by modifying the distribution at run-time (e.g., out-of-distribution examples). **Therefore, this evaluation may generalize to ADS deployments in which models are subject to selection bias, but may generalize less readily to decision-dependent outcome tasks or those with pronounced outcome measurement error.**

A.2.4 Study 4: Zhang et al. [102]. This study examines whether showing model confidence scores (probability estimates) and local explanations helps humans make more accurate decisions while using predictive models. This study also examines whether these decision-time interventions help humans better calibrate trust in the models predictions, defined as following recommendations more often when the model is more confident. To test this hypothesis, the authors trained a model to predict whether an individuals income would exceed $\$50K$ given tabular demographic and job information from the UCI Adult Data Set. We show the sub-region of focus and ADS regime in Figure 4.4.

- **Sub-region of focus:** D, X, Y. This study focuses on the sub-region with joint decisions D, observables X, and outcome proxies Y. We include model observables (X) because both the human and the model had access to the same set of 8 attributes about individuals while predicting their income. We include the proxy label (Y) because the authors list a target outcome involving whether an individual makes more or less than $\$50K$. We include joint human model decisions (D) because the experimental treatment involves different decision-time interventions shown to participants (i.e., model confidence scores or explanations). We do not include Y^* because the authors do not examine sources of measurement error that can impact the reported income available in data. The authors leverage the UCI Adult dataset based on 1994 Census Data, which could be subject to various sources of reporting bias. We do not include the edge connecting decisions D and outcomes Y, Y^* because participant guesses do not influence the income of individuals in the dataset.
- **ADS Regime: decision-independent target variable.** Because participant responses do not influence the income of participants, this task includes decision-independent target variables. As a result, confounding, selection bias, intervention effects, and selective labels are not a concern in this task. **As a result, findings from this study may be most likely to generalize to other decision-independent target variable tasks (e.g., house price prediction, jellybean counting) and may be less likely to generalize to real-world predictive model deployments with decision-dependent outcomes.**

A.2.5 Study 5: Gordon et al. [41]. This work proposes a normative and technical framework called Jury Learning, which is intended to help practitioners *"recognize and integrate annotator disagreement in the classifier pipeline [41]."* Under the proposed framework, model developers specify groups of users whose opinions should be considered during moderation decisions (i.e., juries), along with a relative weighting of each group. At inference time, a model predicts the annotations of each individual annotator, and a final decision is reached by combining predictions via the specified jury rule.[5] As part of the framework evaluation, the authors recruited online moderators from Discord, Twitch, and Reddit, and evaluated the diversity of annotator pools constructed via Jury Learning against a baseline of "majority vote" aggregation in a comment toxicity classification task. Thus, this study differs from those discussed above because the involvement of human subjects occurs at *model development time* rather than at *decision time*. Nevertheless, this toxicity classification task falls within our framework scope (§ 3.1). We show the sub-region of focus and ADS regime in Figure 4.5.

- **Sub-region of focus:** X, Y, Y^*. We include model observables (X) because both the human and the model have access to the same set of information about comments. We include Y and Y^* because the study investigates how practitioners construct differing jury rules for mapping observed ratings from participants (Y) to the latent construct of "toxicity" being predicted by the model (Y^*). The D region is not included in this study because the authors do not examine content moderation decisions or toxicity ratings at *deployment time*. We omit unobservables Z because the authors do not study how unobserved information could impact toxicity perceptions of annotators or the learned jury decisions.
- **ADS Regime: decision-independent.** In this setting, the label targeted by toxicity classification models is determined by the subjective opinion of the annotator viewing the content. As a result, measurement error is relevant in this setting because the operationalization of "toxicity" targeted by the model depends on the identity of the user, the context in which the post is viewed, and the annotation protocol, among other factors. However, confounding, selection bias,

[5] See [41] for framework details not discussed in this summary, such as repeated sampling over several trials.

selective labels, and treatment effects are not of concern in this setting because there is not a time dependency of decisions and outcomes. **Therefore, findings from this study may be most likely to generalize to other decision-independent target variable tasks (e.g., house price prediction, jellybean counting) and may be less likely to generalize to real-world predictive model deployments with decision-dependent outcomes.**

A.2.6 Study 6: Green and Chen [42]. This work examines whether risk assessments improve the accuracy, fairness, and reliability of human decisions in financial lending and recidivism prediction tasks. The authors train risk assessments to predict re-arrest and loan default outcomes given tabular administrative data. The experimental conditions test several variations of the procedure for presenting risk assessment information to participants (e.g., no score, local explanation, immediate outcome feedback) before participants make the final decision. We show the sub-region of focus and ADS regime in Figure 4.6.

- **Sub-region of focus:** D**,** X**,** Y. We include the D region because experimental conditions manipulate the joint human-model decision-making process. We include the X region because participants were provided with a narrative profile containing factual content that coincides with the model training features (e.g., defendant age, applicant credit score). We include the proxy label region Y because the authors list a target outcome consisting of failure to appear or re-arrest (recidivism) and loan default. We omit the target variable Y^* because the authors do not examine sources of measurement error impacting recorded re-arrest and default outcomes (e.g., crimes that go unreported). We do not bold the arrow from D to Y and Y^* because the authors do not examine how the historical decisions of judges or loan officers might influence the outcomes available for the applicant pool.
- **ADS Regime: decision-dependent target variable.** Both experimental tasks included in this study involve a setting in which a model is trained on data from decisions made under an earlier decision-making policy. As a result, loan repayment is only observed among approved applicants, while re-arrest and failure to appear is only observed among released defendants. As a result, the model included in this experimental task is subject to selection bias, selective labels, confounding, intervention effects, and measurement error. **Therefore, findings from this study may be most likely to generalize to other decision-dependent target variable tasks.**