Under-reliance or misalignment? How proxy outcomes limit measurement of appropriate reliance in AI-assisted decision-making

LUKE GUERDAN, Carnegie Mellon University, USA
KENNETH HOLSTEIN, Carnegie Mellon University, USA
ZHIWEI STEVEN WU, Carnegie Mellon University, USA

As AI-based decision support (ADS) tools are broadly adopted, it is critical to understand how humans can effectively incorporate AI recommendations into their decision-making. However, existing research studying how humans calibrate their reliance on AI recommendations often overlooks a key difference between human and AI judgments. Whereas humans reason about the broader phenomena of interest in a decision — for example, creditworthiness, disease status, or recidivism risk — AI models predict narrow outcomes based on readily available historical data. Because these observed outcomes are merely proxies of the broader phenomena considered by decision-makers, they may be subject to various forms of bias. We refer to this gap between human and AI decision-making goals as *outcome measurement error*. We argue that failing to address outcome measurement error can produce misleading evaluations of "appropriate reliance" in AI-assisted decision-making. In particular, we identify three sources of outcome measurement error that can bias naive evaluations of appropriate reliance. Based on a broad synthesis of existing literature, we propose a unifying framework for describing outcome measurement assumptions. We use our framework to identify future lines of research that account for outcome measurement in the evaluation of appropriate reliance.

Additional Key Words and Phrases: AI-assisted decision-making; Measurement; Artificial intelligence; Trust

ACM Reference Format:
Luke Guerdan, Kenneth Holstein, and Zhiwei Steven Wu. 2022. Under-reliance or misalignment? How proxy outcomes limit measurement of appropriate reliance in AI-assisted decision-making. 1, 1 (April 2022), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

AI-based systems are increasingly being used to augment human judgement in domains such as healthcare, lending, and child welfare. As humans integrate AI predictions and recommendations into their decision-making, a growing body of work has focused on achieving *complementary performance*, where the human-AI team makes better decisions than either acting in isolation [3, 33]. To achieve such complementary performance, humans must know to rely on AI-based recommendations when appropriate, but to follow their own judgement otherwise. By establishing measures of *appropriate reliance*, it may be possible to identify cases when humans are blindly following recommendations (i.e., *over-reliance*) or disregarding recommendations too often (i.e., *under-reliance*) [27].

However, a central limitation of current evaluations of *appropriate reliance* in AI-assisted decision-making is that they are performed with respect to the outcomes considered by AI models, not necessarily those of interest to the human decision-makers within human-AI teams. Work assessing the quality of human-AI decision-making often reports results

Authors' addresses: Luke Guerdan, lguerdan@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, USA; Kenneth Holstein, Carnegie Mellon University, Pittsburgh, USA, kjholste@cs.cmu.edu; Zhiwei Steven Wu, Carnegie Mellon University, Pittsburgh, USA, zstevenwu@cmu.edu.

in terms of *"performance"* operationalized by accuracy, AUC, or a similar statistical measure [3, 24, 25]. Researchers claim humans are under-reliant when they fail to follow the AI's recommendation despite it being *"correct"*. Similarly, researchers report over-reliance in cases where humans follow the model despite it being *"incorrect"* [4, 8]. Yet these notions of decision quality hinge on the definition of ground truth considered by the AI system. Consequentially, resulting assessments of human reliance are only valid to the extent that the ground truth encoded in labels accurately reflects the goals of decision-makers.

In contrast to the observable proxy outcomes typically considered by AI systems, humans make decisions based on meaning-rich representations of the world, including latent constructs [17, 23]. For instance, a loan officer might consider the creditworthiness of a lendee while deciding whether to approve a new line of credit; or a physician might evaluate cardiovascular disease risk holistically as a combination of a number of factors. "Creditworthiness" and "cardiovascular disease risk" are examples of latent constructs that are not directly observable in the world, but can be operationalized via a measurement model [21]. AI-based decision support tools assume de-facto measurement models when they assign an outcome label – loan default or heart attack events, for example – as a proxy for the latent construct of interest. We refer to this gap between the construct of interest to human decision-makers and the proxy used by AI systems as *outcome measurement error*.

Outcome measurement error shapes expert reliance in real-world AI-assisted decision-making deployments. In pre-trial detention decisions, juvenile defendants are commonly identified as high-risk by AI systems due to the connection between youth and re-arrest [15]. However, judges sometimes override these high-risk scores because youth is seen as a mitigating factor for culpability [32]. Here, the gap between judges' notion of "risk to society" and the re-arrest proxy targeted by the AI contributes to apparent "under-reliance". In child welfare, AI-based decision support tools have been introduced to help social workers assess whether children are at risk of abuse. One tool, the Allegheny County Family Screening Tool (AFST), uses placement in foster care within two years as a proxy for abuse and neglect. Yet social workers have reported overriding the AFST's recommendations, in part because its long-term view of risk and its focus on predicting *placement in foster care* is incongruent with their own focus on assessing actual, near-term safety risks to children [22].

Based on the challenges discussed above, there is a pressing need to develop strategies for evaluating appropriate reliance in the presence of outcome measurement error. However, a key impediment to establishing these strategies is that the sources and implications of outcome measurement error remain poorly understood in AI-assisted decision-making. Our current understanding of outcome measurement error is scattered across work in quantitative social sciences [21, 29], economics [6, 23, 30], and machine learning [11, 12, 26], with limited synthesis of general trends and best practices. At a basic level, we lack a synthesis of the common error sources that should be considered while developing measures of appropriate reliance. Moreover, we do not have a framework for articulating outcome measurement assumptions, along with their implications on assessing appropriate reliance. This makes it difficult to understand how existing assessments of appropriate reliance might be impacted by outcome measurement error.

Therefore, in this work, we provide a foundation for characterizing sources and implications of outcome measurement error in AI-assisted decision-making, and use this framework to highlight nuances in the evaluation of appropriate reliance. Our work includes two key contributions. **First, we identify outcome measurement-related challenges that should be considered while developing assessments of appropriate reliance in AI-assisted decision support** (Section 2). In particular, we highlight challenges arising from the datasets used to develop AI-based decision support tools, and identify three sources of outcome measurement error. We discuss how evaluations of appropriate reliance could be incomplete if these challenges are not appropriately addressed. **Second, we provide a unifying**

framework for characterizing the outcome measurement process in AI-assisted decision support (Section 3). We propose a three-step process — measurement, prediction, and evaluation — involved in the development of AI-based decision support tools. We use this framework to describe a broad set of AI-based decision support methods that have been proposed in the literature (see Table 1).

2 BARRIERS TO MEASURING APPROPRIATE RELIANCE IN AI-ASSISTED DECISION-MAKING

In this section, we draw attention to challenges that, left unaddressed, may lead to misleading evaluations of appropriate reliance. We begin by discussing the datasets used by AI-based decision support tools. We note important distinctions between datasets used to train "real-world" deployments of AI-based systems, and those often included in current evaluations of appropriate reliance [3, 4, 8, 35]. Based on these differences, we then highlight three forms of outcome measurement error — construct validity of outcome proxies, intervention effects, and selective labels – that are likely to impact assessments of appropriate reliance.

2.1 Datasets used to develop AI-assisted decision-making systems

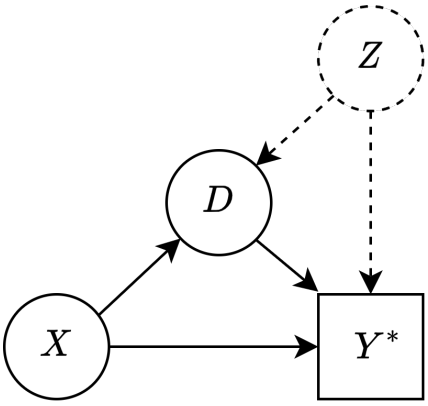

Fig. 1. Structural causal model showing the data generative process that gives rise to training data used by AI-based decision support tools. X represents observable data that plays a factor in the expert decision, Z represents unobserved information only available to humans, D represents historical expert decisions, and Y^* represents the primary outcome of interest to decision-makers. We show Y^* in a box rather than a circle to highlight that this outcome is operationalized via a measurement model rather than observed directly.

One barrier to evaluation of appropriate reliance is that datasets used to examine reliance patterns in lab-based studies differ from those used to train real-world AI-based decision support tools. Lab-based studies often conduct empirical evaluations using simplified experimental tasks to make the evaluation feasible with online participants who lack domain expertise [3, 4, 8, 35]. For instance, previous reliance evaluations have asked participants to predict the sale price of houses [8], the nutrient content of food [4], or forest cover from images [35]. However, datasets used in these tasks differ from real-world decision support settings in important ways.

In real-world decision support settings, the AI-based tool is intended to improve existing decision-making practices. Therefore, model developers rely on historical data from past human decisions to train the model embedded in the tool. Using data generated from past human decisions introduces important nuances that should be considered during model development *and* during assessment of appropriate reliance. For instance, in the course of making their decisions,

humans might have had access to additional information that is not available within the recorded dataset. Additionally, the human decision may have *influenced* the outcome observed in the historical data. In contrast, datasets used in many lab-based assessments of reliance are not subject to the influence of past human decisions (an exception includes recidivism prediction in [35]).

To improve the ecological validity of lab-based assessments of appropriate reliance, it is critical to understand the real-world processes that give rise to the datasets upon which AI-based decision support tools are trained. In causal inference and statistics, these real-world processes are often described by *data generative processes*. Causal directed acyclic graphs (DAGs) provide one graphical tool for reasoning about a data generative processes, where arrows reflect causal relationships between different nodes [31]. We propose a DAG that describes the data generative process for datasets used by AI-based decision support tools in Figure 1. In this figure, X involves observed information (e.g., medical history, public welfare records) that is digitally recorded and leveraged as features in a model. Additionally, Z represents contextual information that is available to experts but unavailable to a model. Z is sometimes referred to as a confounder or unobservable in AI-assisted decision-making literature (where confounders are typically indicated by dotted lines in DAGs) [12, 23]. In the figure, D describes the decision reached by the human based on X and Z. Finally, Y^* shows the historical outcome that resulted from X, Z, and D.

The outcome Y^* shown in Figure 1 plays a central role in the expert decision-making process. Specifically, Y^* represents the unobserved outcome of interest that humans consider as they weigh their decision. A judge might consider Y^* to be "violent crime", while a social worker might consider Y^* as "serious abuse and neglect". However, because Y^* is operationalized via a measurement model rather than observed directly, we represent Y^* via a square in Figure 1. In practice, the most commonly used measurement model is to simply use an observed outcome (e.g., re-arrest, placement in foster care) as a proxy for the construct of interest. More nuanced measurement models have also been proposed [12, 29]. However, a barrier to establishing measures of appropriate reliance is that the outcome measurement process is poorly understood. This poses a challenge in light of the issues we discuss below.

2.2 Outcome measurement error

When AI-based decision-support tools are developed using data generated via Figure 1, this introduces several key issues in the estimation of Y^*. First, because the true outcome of interest in AI-assisted decision settings is unobservable, there can be a gap between Y^* and the proxy adopted by the measurement model (Y). In these settings, it is critical to consider the size of this gap by examining the *construct validity of outcome proxies*. An additional impediment to predicting Y^* arises from the results of past human decisions. As highlighted in Figure 1, human decisions are not passive, independent events. Instead, they act as *interventions* that change the probability of Y^* (i.e., *intervention effects*). For example, an opened child welfare investigation might connect a family with resources that reduces the chances of child abuse. A related barrier to estimating Y^* is that human decisions change the probability of observing an outcome proxy. For example, reoffence is only observed by released defendants, and defaults are only observed by borrowers who receive a loan. Due to these *selective labels*, historical decisions influence the outcomes available for modeling. We refer to this collective set of issues in estimating Y^* via observed outcome proxies Y as *outcome measurement error*.

Despite its importance in establishing measures of appropriate reliance, outcome measurement error remains largely unacknowledged in AI-assisted decision-making literature [2–5, 8, 17, 24, 25, 35]. Instead, evaluations assume that the outcome proxy predicted by the AI-based tool (Y) is equivalent to the true outcome of interest (Y^*). This poses an issue in light of the challenges raised above. Specifically, because existing measures of appropriate reliance hinge on the definition of a "correct" decision by a human or AI tool, and "correct" outcomes are determined based on proxy

outcomes, this could potentially lead to misleading evaluations. As a result, measures of appropriate reliance developed in lab-based studies with limited outcome measurement error may fail to generalize when deployed in real world settings where the gap between Y and Y^* is more pronounced. We now discuss three sources of outcome measurement error in detail, and highlight how they can lead to misleading assessments of reliance.

2.2.1 Construct validity of outcome proxies. One source of outcome measurement error that plays a key role in establishing assessments of appropriate reliance is the *limited construct validity of outcome proxies*. Construct validity broadly describes how well latent phenomena of interest to humans is operationalized by a measurement model [21]. One key sub-component of construct validity especially relevant to AI-based decision support is content validity, which assesses the degree to which an operationalization fully and completely captures the unobserved phenomena of interest. In lending, a loan default proxy (Y) might demonstrate poor content validity for creditworthiness (Y^*) in cases where default occurs due to unforeseen external events (e.g., viruses, natural disasters) unrelated to responsible fiscal behavior. In pretrial risk assessments such as COMPAS, a re-arrest proxy (Y) might demonstrate poor content validity for “risk to society" (Y^*) when unwarranted arrests are made or when committed crimes go undetected [15].

Failing to consider the construct validity of outcome proxies can lead to misleading evaluations of appropriate reliance. In particular, **in cases where construct validity of outcome proxies is poor, assessing appropriate reliance with respect to the proxy instead of the outcome of interest to decision-makers can lead to mistaken conclusions that humans are under-relying on the AI model**. For example, because judges consider “risk to society" (Y^*) on the basis of the likelihood of reoffence as well as culpability, they are more likely to release young defendants [32]. These young children might be considered as high-risk when evaluated on the basis of a re-arrest proxy (Y) alone. If an AI-based decision support tool predicts Y without taking this difference into consideration, it may observe many judge over-rides in cases with young defendants. Under current definitions of under-reliance (i.e., human overrides when an AI-based tool is “correct"), this behavior from the judge will be seen as undesirable. As a result, interventions targeted at mitigating “under-reliance" might be designed to encourage the judge to disregard their own judgement in favor of the more narrow outcome definition adopted by the AI-based decision support tool.

When the construct validity of outcome proxies is in doubt, this presents a systemic challenge to the assessment of appropriate reliance. To understand this challenge, we can consider possible relationships between the outcome of interest (Y^*), the proxy adopted by the AI-based decision support tool (Y), and historical human decisions (D) that may occur in an AI-based decision-support context. We depict this relationship via a tree-way Venn diagram in Figure 2, and consider the implications of each region in turn:

- **R1.** The first region shows instances where the human and outcome proxy both fail to capture Y^*. In an AI-based decision support context, this might occur when a child is not screened-in ($D = 0$), not placed within two years ($Y = 0$) but suffers serious abuse and neglect ($Y^* = 1$). Left unaddressed, this will cause the human-AI team to systematically miss instances of the outcome of interest to decision-makers. Previous work on AI-assisted decision-making has referred to this region as omitted payoff bias [12, 23].
- **R2.** The second region contains instances of Y^* that are captured by the definition of the proxy outcome, but not captured by historical human decisions. AI-based decision support tools may be particularly effective at identifying instances within this region. Therefore, while developing measures of appropriate reliance, it may be beneficial to encourage humans to trust the judgement of AI systems for fore instances within this region.

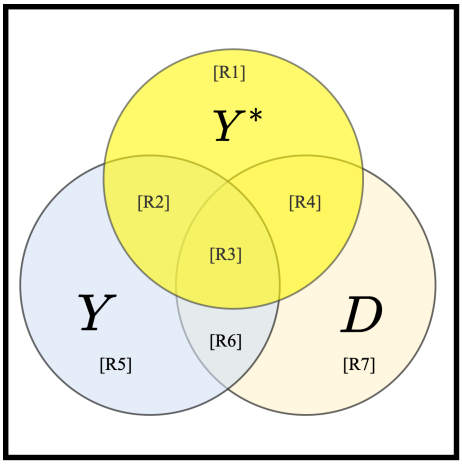

Fig. 2. Venn diagram showing possible combinations of $Y = 1$, $Y^* = 1$, and $D = 1$ when Y is an imperfect proxy of Y^*. Each region shows instances of human-AI decision-making that can occur in a given AI-based decision support context.

- **R3.** The third region contains clear instances of Y^* that are captured by both the observed outcome and the human. Existing measurements of appropriate reliance may be more likely to correctly assess instances within this region as times when human judgement is well-calibrated with the AI's recommendation.
- **R4.** The fourth region contains instances of Y^* that are captured by human decisions but not by the outcome proxy. Critically, instances within this region may be *incorrectly classified as under-reliance* under current evaluations. This is because humans correctly identify Y^*, but the proxy targeted by the AI-based tool does not. Corrected measures of appropriate reliance should take this region into consideration in order to accurately reflect the goals of human decision-makers.
- **R5-7.** Regions 5-7 show instances where the outcome label, the human, or both, are false positives with respect to Y^*. R6 shows cases where Y and D are in agreement, but both fail to detect Y^*. Current evaluations may mistakenly interpret instances in this region as *appropriate reliance* when a model predicting Y and the human D are in agreement (i.e. $\hat{Y} = 1$, $D = 1$, and $Y = 1$). Further, evaluations may mistakenly interpret instances in this region as *over-reliance* when a model predicting Y and D are in disagreement (i.e. $\hat{Y} = 0$, $D = 1$, and $Y = 1$). However, both of these conclusions would be mistaken given that the true outcome of interest Y^* did not occur.

In addition to construct validity of outcome proxies, there are also other outcome measurement challenges that should be considered while assessing reliance in AI-assisted decision-making.

2.2.2 Intervention effects. Another source of outcome measurement error that impacts the assessment of appropriate reliance is *intervention effects* resulting from past human decisions [11]. To understand this concern, we return to the data generative process discussed in Section 2.1. Note that historical human decisions constitute an intervention on Y^*, and therefore change the probability of observing its proxy Y in historical data. For example, consider a child maltreatment setting, whereby social workers aim to prevent abuse via effective allocation of welfare resources. Here, the goal of social workers is not simply to detect maltreatment cases via accurate decisions. Rather, call workers aim to make it less likely that a child will be seriously harmed through interventions that connect the family with appropriate

support [22]. Whereas humans consider the question "How likely is Y^* under the proposed decision?", AI-based tools trained on data generated via Figure 1 answer the question "How likely is Y^* under historical human decisions?".

This difference between how humans and AI-based decision support tools reason about Y^* complicates the assessment of appropriate reliance. This is because models trained on data generated via Figure 1 *systematically underestimate the risk of cases that benefited most from the decision* [11]. For example, in the child maltreatment context, if a social worker elected to open an investigation on a high-risk family, and that investigation averted abuse (Y^*), the data available for modeling will show the case data (X) along with a positive outcome (Y) (i.e., no removal from the home in two years). In this case, a human deciding whether to follow the AI-tool's recommendation *should follow their own judgement and open an investigation* even though this is contrary to the outcome proxy. If a measure of appropriate reliance does not fully capture this nuance, it may incorrectly infer that the human is "under-reliant" on the model when they decide to intervene despite a low-risk prediction.

2.2.3 Selective labels. A final barrier we highlight to assessing appropriate reliance in AI-assisted decision-making is *selective labels* [23, 26]. Selective labels describe a setting where outcome proxies are only observed among cases where humans decided to intervene. For instance, in a pretrial setting, we only observe re-arrest (Y) in cases where judges decided to release the defendant. In lending settings, we only observe default (Y) in cases where a loan was approved by the lender. Initial work on selective labels highlights the importance of considering selective labels while evaluating appropriate reliance. In particular, previous work has shown that after accounting for the effects of selective labels, human decisions are more predictive of proxy outcomes than they would be under a naive assessment that doesn't take selective labels into account [23, 26]. Therefore, accounting for selective labels is an important step to developing a suitable baseline measure of human vs. AI decision quality.

3 MEASUREMENT, PREDICTION, AND EVALUATION: A UNIFYING FRAMEWORK

In the previous section, we discussed outcome measurement challenges in AI-assisted decision-making and their relation to assessments of appropriate reliance. This raises the question: "how should assessments of appropriate reliance account for outcome measurement error?" This question is difficult because accounting for outcome measurement error requires understanding the construct of interest to decision-makers. For instance, in the Venn diagram presented in Figure 2, it is challenging to establish which region a particular case occupies without understanding how D, Y, and Y^* are interrelated. Initial work in AI-assisted decision-making has proposed solutions to some aspects of outcome measurement error [11, 12, 26]. However, community members establishing measures of appropriate reliance may not be fully aware of this work. This is understandable, as discussion of outcome measurement error remains scattered among the quantitative social sciences, economics, and machine learning communities (Table 1). Therefore, drawing from this literature, we provide a unifying framework for characterizing outcome measurement in AI-assisted decision-making. Our framework captures a broad set of previously proposed modeling approaches and contains three aspects: *measurement*, *prediction*, and *evaluation*.

3.1 Measurement

During the measurement step of model development, the unobserved outcome of interest to the organization (Y^*) is estimated using observations from the data generative process in Figure 1. This training data can be described by the tuple (X, Z, D, W), where each element is a random variable involved in the data generative process and W is a set of observed outcome proxies $W = \{Y^1, \ldots, Y^k\}$. Here, we use W instead of Y to represent the more general case where

multiple outcome proxies are used in parallel to approximate Y^*. In a general form, the measurement model used in AI-assisted decision-making is some function F_m of the data (X, Z, D, W) available to the organization. Given this data, the organization develops an estimate for Y^* via:

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

We highlight two key aspects of this measurement model. First, the organization has full access to (X, Z, D, W) based on historical decisions made prior to the introduction of the decision support tool. Recall from Figure 1 that Z is often unobserved by the organization. For full generality, we include Z in this definition because existing work also considers settings where unobservables are available during tool development, but not during deployment [10].

Establishing F_m involves *outcome measurement assumptions* made by the tool developers and organization. There is no direct way to estimate Y^* without measurement assumptions because Y^* is not directly observed in historical data. For example, one common assumption adopted in practice is that an observed outcome proxy is a reliable estimate for Y^*. In these cases, F_m can be given by $\hat{Y}^* = Y^1$. Here, $Y^1 \in W$ is a single observed outcome proxy (referred to as Y in Section 2). Though this assumption is most common, others have been adopted in AI-assisted decision-making and quantitative social sciences literature. For example, a broad class of work in quantitative social sciences uses Latent Class Analysis (LCA) to estimateY^* based on a set of multiple observed proxies W [29]. This method makes the measurement assumption that $\{Y^1, Y^2, \ldots, Y^K\} \in W$ are conditionally independent given Y^*. As an additional example, a recently-proposed AI-assisted decision-making approach learns a measurement model F_m by assigning $Y^* = 1$ when experts are predicted to agree, and uses a proxy Y^1 otherwise [12]. This work makes the measurement assumption that expert decision consistency will be more likely when $Y^* = 1$. Table 1 builds on these examples by describing the measurement model (F_m) assumed by a broader set of AI-based decision support work.

3.2 Prediction

After establishing a measurement model to estimate Y^* given (X, Z, D, W), organizations and researchers then develop a *prediction model* for use in decision-support settings. This prediction model takes observable features about an individual (X) and makes a prediction about the unobserved outcome of interest (Y^*) established during the preceding measurement step. Because W and Z are unavailable during deployment of the AI-based decision support tool, these are not included in the prediction model. Typical AI-based decision support workflows do not assume that human decisions are available at run-time. This is because the algorithmic recommendation is designed to *augment rather than replace* the human decision-maker (i.e., algorithm-in-the-loop) [2, 3]. Nevertheless, a set of recently proposed AI-based decision support methods also consider the case where a human decision is available at runtime, and available to the model (i.e., human-in-the-loop) [16, 28, 33, 36]. Therefore, to capture the full generality of possible prediction models, we include both X and D as possible prediction inputs. Given X and optionally D available at runtime, the prediction model estimates:

$$\hat{Y} = F_p[\hat{Y}^* | X, D] \tag{2}$$

Note that this step mirrors current AI-based decision support tools that predict an estimate $\hat{Y}$ for the observed outcome proxy Y^1. The key distinction we draw here is that this prediction model is actually estimating a measurement model $\hat{Y}^*$ established during the outcome measurement step. Table 1 shows the prediction model (F_p) used by a broad set of AI-based decision support methods.

Work	Measurement (F_m)	Prediction (F_p)	Evaluation	Challenge(s)
Gao et al. [16]	$\hat{Y}^* = F_m[W]$, where $W = Y^1$. Assumes that outcome proxy is ground truth.	D available at run-time (human-in-the-loop). Work improves F_p using human decisions.	No assessment of F_m. Accuracy and ROC based evaluation of F_p.	None
Madras et al. [28]				
Wilder et al. [36]				
Tan et al. [33]				
Hilgard et al. [20]		D un-available at runtime (algorithm-in-the-loop)	Accuracy-based evaluation of F_p with focus on human decision performance. No evaluation of F_m.	
De-Arteaga et al. [12]	$\hat{Y}^* = F_m[W, D, X]$ with $W = \{Y^1\}$. Assumes Y^* can be identified via consistancy of expert decisions.		Assess F_p via precision/recall of on training outcome Y^1 and F_m via precision/recall on held-out outcomes Y^2, Y^3.	Construct validity, proxy observation bias
Coston et al. [11]	$\hat{Y}^* = F_m[W, D]$, where $W = \{Y^1\}$ and Y^1 only observed when $D = 1$.		F_m uses doubly-robust estimation to account for treatment effects of D.	Intervention effects
Lakkaraju et al. [26]	$\hat{Y}^* = F_m[W, D]$, where $W = \{Y^1\}$ and Y^1 is observed when $D = 1$.		Proposes method for accounting for selective labels during the evaluation of F_p	Proxy observation bias
Kleinberg et al. [23]	$\hat{Y}^* = F_m[W]$, where $W = Y^1$ and Y^1 is only observed when $D = 1$.		Uses contraction method proposed in [26]. Evaluates model on a set of held-out outcomes	Construct validity, Proxy observation bias
Coston et al. [10]	$\hat{Y}^* = F_m[X, Z, D, W]$, where $W = \{Y^1\}$. Assumes confounders Z available at at training.		F_m uses doubly-robust estimation to account for treatment effects of D.	Intervention effects
Latent Class Analysis (McCutcheon [29])	$\hat{Y}^* = F_m[W]$. Assumes $\{Y^1, ..., Y^K\}$ are conditionally independent given latent class membership.	3-step LCA with covariates [34]. Could include settings where D is known or unknown at runtime.	Assess F_m with G^2 or BIC. F_p fit with a logit model.	Construct validity
Fogliato et al. [14]	$\hat{Y}^* = F_m[W]$, where $W = Y^1$. Assumes measurement error between Y^* and Y^1 with known magnitude.	N/A	Primary contribution is a statistical framework for assessing measurement error (termed *target variable bias*).	Construct validity

Table 1. Taxonomy of measurement, prediction and evaluation challenge(s) addressed by a broad set AI-assisted decision-making methods drawn from risk assessment and algorithmic decision support literature. We show cases where an outcome proxy is assigned as the measurement hypothesis by $\hat{Y}^* = F_m[W]$, where $W = Y^1$.

3.3 Evaluation

A final stage of AI-based decision support tool development involves evaluation. In classical modeling contexts, evaluation efforts assess F_p in terms of accuracy, AUC, or statistical fairness measures with respect to a *proxy outcome*. In addition to model-based evaluations, it is becoming increasingly common to assess how the system is adopted by humans in practice [1, 23]. For example, an evaluation might assess whether decision-makers "over-rely" or "under-rely" on the tool [9]. Because these evaluations are with respect to F_p, they are suitable in cases where the prediction is aligned with the interests of the decision-maker, but incorrect in a specific instance (e.g., due to missing contextual information or dataset shifts). However, this evaluation of F_p is *distinct from an evaluation of the measurement process* F_m that determines whether the model effectively targets the construct of interest Y^*.

Recently proposed statistical methods propose some avenues for evaluating F_m at a model level [11, 14, 23, 26]. At a broader level, tools used to assess construct validity are also used to assess different aspects of F_m. For example, organizations might examine the convergent validity of F_m to see whether estimates of Y^* correlate with other known measurements of the same unobserved outcome [21]. They might also examine the predictive validity by checking whether estimates of Y^* are predictive of downstream external outcomes. Measurement theory in the quantitative social sciences provides a rich set of ways to evaluate the construct validity of [19]. Notably, however, current measurements of over-reliance and under-reliance do not involve F_m, which makes it difficult to assess how the measurement model will impact resulting human trust and decision-making. We summarize evaluation-related procedures adopted by a broad set of AI-assisted decision-making literature in Table 1.

4 GUIDELINES AND IMPLICATIONS

In the discussion above, we described several outcome measurement-related challenges that may limit existing assessments of appropriate reliance. In light of this discussion, we now provide suggestions for the development of appropriate reliance measures in AI-assisted decision support. Most broadly, the research community should consider outcome measurement error carefully during the development of reliance measures. This is because if an evaluation of reliance mistakenly claims that a human is *under-relying* on a prediction model (F_p) when they are in fact disagreeing with the outcome measurement (F_m), organizations may pressure decision-makers to follow the guidance of misaligned tools. Applying excessive organizational pressure to follow incorrect tools has the potential to introduce damaging consequences [7, 13, 18, 22].

A first step that the community can take to better address outcome measurement error in assessments of reliance is to consider a more realistic suite of datasets in empirical evaluations. Ideally, these datasets should involve data drawn from historical human decision-making processes (Figure 1), as these better reflect real-world deployments of AI-assisted decision-making systems. As a second step forward, we suggest developing mechanisms for assessing whether human over-rides of AI recommendations are indicative of under-reliance (F_p), or are symptoms of a more fundamental outcome measurement issue (F_m). Further, it may also be valuable to investigate how closely humans scrutinize an AI's recommendation under different outcome measurement conditions. For example, when the proxy targeted by a model differs from the construct of interest, humans may disengage and *over-rely* on the model recommendations.

Finally, measures of appropriate reliance should be developed that account for nuanced issues such as intervention effects and selective labels. Because of the complexity of these issues, and the unknown effects they may have on human decision-making, it may be necessary to conduct controlled empirical experiments to understand their effect on human reliance patterns. We hope that our framework for characterizing outcome measurement error through the lens of measurement, prediction, and evaluation will provide a valuable springboard for these research efforts.

REFERENCES

[1] Alex Albright. 2019. If you give a judge a risk score: evidence from Kentucky bail decisions. (2019).

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Cheng and Stapleton, 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)*.

[8] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[9] 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.

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

[11] 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.

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

[13] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[14] 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.

[15] 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.

[16] 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).

[17] 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.

[18] 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.

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

[20] 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.

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

[22] Anna Kawakami, Venkat Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yang Cheng, Diana Qing, Adam Perer, Steven Wu, Zhu Haiyi, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support.. In *To appear in Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI'22)*.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[28] David Madras, Toniann Pitassi, and Richard Zemel. 2018. Predict responsibly: Increasing fairness by learning to defer. (2018).
[29] Allan L McCutcheon. 1987. *Latent class analysis*. Number 64. Sage.
[30] Sendhil Mullainathan and Ziad Obermeyer. 2017. Does machine learning automate moral hazard and error? *American Economic Review* 107, 5 (2017), 476–80.
[31] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[32] Megan T Stevenson and Jennifer L Doleac. 2021. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440* (2021).
[33] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).
[34] Jeroen K Vermunt. 2010. Latent class modeling with covariates: Two improved three-step approaches. *Political analysis* 18, 4 (2010), 450–469.
[35] 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.
[36] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).