A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms

Amanda Coston
Heinz College & Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
acoston@cs.cmu.edu

Anna Kawakami
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, USA
akawakam@andrew.cmu.edu

Haiyi Zhu
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, USA
haiyiz@cs.cmu.edu

Ken Holstein
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, USA
kjholste@cs.cmu.edu

Hoda Heidari
Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
hheidari@cs.cmu.edu

***Abstract*—Recent research increasingly brings to question the appropriateness of using predictive tools in complex, real-world tasks. While a growing body of work has explored ways to improve value alignment in these tools, comparatively less work has centered concerns around the fundamental justifiability of using these tools. This work seeks to center validity considerations in deliberations around whether and how to build data-driven algorithms in high-stakes domains. Toward this end, we translate key concepts from validity theory to predictive algorithms. We apply the lens of validity to re-examine common challenges in problem formulation and data issues that jeopardize the justifiability of using predictive algorithms and connect these challenges to the social science discourse around validity. Our interdisciplinary exposition clarifies how these concepts apply to algorithmic decision making contexts. We demonstrate how these validity considerations could distill into a series of high-level questions intended to promote and document reflections on the legitimacy of the predictive task and the suitability of the data.**

***Index Terms*—predictive analytics, validity, deliberation, algorithmic oversight, responsible AI, algorithmic decision support**

I. Introduction

Data-driven algorithmic decision-making, in theory, can afford improvements in efficiency and the benefits of evidence-based decision making. Yet in practice, data-driven decision systems, often taking the form of algorithmic risk assessments, have caused significant adverse consequences in high-stakes settings. Investigators have identified unintended and often biased behavior in algorithmic decision systems used in a variety of applications, from detecting unemployment and welfare fraud to determining pre-trial release decisions and child welfare screening decisions, as well as in algorithms used to inform medical care and set insurance premiums [1, 2, 3, 4, 5, 6, 7, 8]. These high-profile incidents have brought into focus key questions such as how we can anticipate these harms before deployment, and perhaps more fundamentally, whether algorithms are suitable in the first place for such high-stakes decision-making tasks.

In this work, we examine how validity considerations can help guide decisions about whether to build and deploy algorithmic decision systems. Our proposal can be contextualized in the tradition of technology refusal. Activists have long argued for the value in refusing technology and opting not to build [9, 10]. These calls have taken on new urgency in the modern setting of algorithmic proliferation as many scholars and activists debate when to repair or abolish the use of algorithms in socially consequential settings [11, 12, 13, 14, 15, 16, 17].

To anticipate harms before deployment, researchers and practitioners have proposed a suite of tools and processes. This work has frequently considered questions of value-alignment, such as how to promote fairness and establish transparency and accountability [18, 19, 20, 21, 22]. More recently, there have been growing calls to assess the appropriateness of using predictive tools for complex, real-world tasks from a *validity* perspective [23]. In many cases where algorithms prove unsuitable for real-world use, the problem originates in the initial problem formulation stages [24, 25], or in the process of operationalizing latent constructs of interest (e.g., worker well-being, risk of recidivism, or socioeconomic status) via more readily observable measures and indicators [26, 27, 28]. Without addressing these issues directly, it may be challenging or impossible to align the resulting model with human values after the fact. In some cases, efforts to do so may actually backfire because of unaddressed upstream issues.

Our work seeks to center validity considerations, a crucial criterion for the justified use of algorithmic tools in real-world decision-making [26, 27, 28]. In doing so, we situate our work

This work was generously funded by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745016, support from PwC and from CMU Block Center for Technology and Society Award No. 53680.1.5007718. Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the authors.

at the intersection of research that debates algorithm refusal versus repair and research that develops artifacts for responsible AI/ML. Guided by the goal of delivering an accessible tool to promote deliberation and reflection around validity, we propose a structure for a protocol designed to distill common validity issues into a question-and-answer (Q&A) format.

The main contributions of this paper are as follows:

1) We provide a working taxonomy of criteria for the justified use of algorithms in high-stakes settings. We utilize this taxonomy to illuminate two important principles for substantiating/refuting the use of ML for decision making: validity and reliability (Section II).
2) We use this taxonomy to conduct an interdisciplinary literature review on validity, reliability, and value-alignment (Section III).
3) We connect modern validity theory from the social sciences to common challenges in problem formulation and data issues that jeopardize the validity of predictive algorithms in decision making (Section IV).
4) We demonstrate how this systematization can inform future work by sketching the structure for a protocol to promote deliberation on validity.

Throughout the paper we will discuss validity in the context of several high-stakes settings where predictive algorithms are increasingly used to inform human decisions: pre-trial release in the criminal justice system and screening decisions in the child welfare system. In the criminal justice setting, judges must decide whether to release a defendant before trial based on the likelihood that, if released, the defendant will fail to appear for trial as well as the likelihood the defendant will be arrested for a new crime before trial [29]. For the child welfare screening task, call workers must decide which reports of alleged child abuse or neglect should be screened in for investigation based on an assessment of the likelihood of immediate danger or long-term neglect if no further action is taken [30].

II. A Taxonomy of Criteria for Justified-Use of Data-driven Algorithms

To assess whether the use of data-driven algorithms is adequately justified in a given decision making context, one must account for a wide range of factors. To give structure to this vast array of considerations, we propose a high-level taxonomy—we posit that the justified use of algorithmic tools requires *at minimum* accounting for validity, value-alignment, and reliability. In this section, we offer a precise definition for these terms. Section III offers an overview of existing literature on each of these topics.

a) The rationale for our taxonomy:: To evaluate whether the use of predictive tools is sufficiently justified in a high-stakes decision making domain, at a minimum, we need to answer the following sequence of questions:

- Can we translate (parts of) the decision making task into a prediction problem where both a measure representing the construct we'd like to predict and predictive attributes are available in the observed data?
- If the answer to the above question is affirmative, does the model we train align with stakeholders' values, such as impartiality and non-discrimination?
- Do we understand the longer-term consequences of deploying the model in decision making processes? For example, how might the deployment setting change over time and can the model be reliably utilized under this changing environment?

The above questions motivate our three high-level categories of considerations for justifying/refuting the use of data-driven algorithms in decision making: validity, value alignment, and reliability.

Before we elaborate on our taxonomy, two remarks are in order. First, we emphasize that a formal, comprehensive taxonomy of considerations around justified-use of algorithms is a formidable research question in itself, and the purpose of our taxonomy is limited to structuring our review of the available literature, tools and resources. We make no claims regarding the comprehensiveness of our taxonomy. We refer the interested reader to treatises on the subject including Fjeld et al. [31], Floridi and Cowls [32], Golbin et al. [33]. Additionally, we note that the three categories at the heart of our taxonomy are intimately connected, rather than mutually exclusive.

Validity. Our first category of considerations, validity, aims to establish that the system does what it purports to do. This quality is much harder to satisfy than one might initially think. Consider for instance the task of predicting which criminal defendants are likely to reoffend. Predictive models are often trained using re-arrest outcomes [34]. Whether a model predicting re-arrest actually predicts reoffense is subject to considerable debate, particularly given that a large body of work has established racial disparities in arrests even for crimes which have little differences in prevalence by race [35]. A model that appears accurate with respect to re-arrests may be quite inaccurate with respect to actual crime. More broadly, the notion of validity requires not only that the system has to predict what it purports to predict, but also must achieve acceptable accuracy both within and outside the training environment (in the real-world deployment). These validity criteria are adapted from validity considerations (e.g., *construct validity*, *internal validity*, and *external validity*) that are widely adopted in social sciences, including psychology, psychometrics, and Human-Computer Interaction [36, 37, 38].

Definition 1 (Validity)**.** *A measure, test, or model is valid if it closely reflects or assesses the specific concept/construct that the designer intends to measure [39].*

We say that a predictive algorithm is valid when it predicts the quantity that we think it does, and similarly we say that an audit or assessment is valid when it evaluates the quantity that we would like to audit or assess. Threats to validity can arise as early as the problem formulation stage where decisions about how to operationalize the problem can induce misalignment between what we intend to predict versus what

the model actually predicts [24, 26]. When validity does not hold, it is quite challenging to assess value-alignment—our next category of considerations. In this sense, we claim that validity is a prerequisite for the more commonly discussed values such as fairness.

Value-alignment. Our second category of considerations focuses on the compliance of the system with stakeholders' values.

Definition 2 (Value-alignment)**.** *Value-alignment requires that the goals and behavior of the system comply with collective values of relevant stakeholders and communities [40].*

Relevant stakeholders might include the communities that will impacted by the algorithmic system or the frontline workers who will work with the system. Commonly discussed values include fairness, privacy, transparency, and accountability. Properties like simplicity and interpretability are often desired as a means to ensure these values [41], and within this taxonomy, we include these properties under the broad umbrella of value-alignment.

Reliability. The final set of considerations that we will discuss concern reliability over time and context.

Definition 3 (Reliability)**.** *Reliability is the extent to which the output of a measurement/test/model is* repeatable, consistent, and stable — *when different persons utilize it, on different occasions, under different conditions, with alternative instruments that measure the same thing [39].*

Reliability concerns in part the dynamical nature of systems in the real world. A system that satisfies our previous two criteria at a given snapshot in time may soon after experience a policy, population, or other notable change that may have profound effects on its validity and value-alignment. Threats to reliability include changes in the population characteristics and/or risk profiles (i.e., distribution shift) or strategic behavior in response to the algorithmic model predictions.

We use this taxonomy to structure a literature review of related work in the following section.

III. Literature Review

In this section we conduct a structured literature review of prior work in validity, value-alignment, and reliability.

A. Validity

We begin our literature review with validity. The machine learning literature has vibrant communities addressing validity-related considerations, such as selection bias and representation bias, but, to the best of our knowledge, there is no unifying validity framework around these issues. For this we turn to the theory of validity in the social sciences. In this section we review key concepts from social science research on validity, and in subsequent sections we translate these concepts to the setting of data-driven algorithms.

Construct validity is concerned with whether the measure captures what the researcher intended to measure. Modern

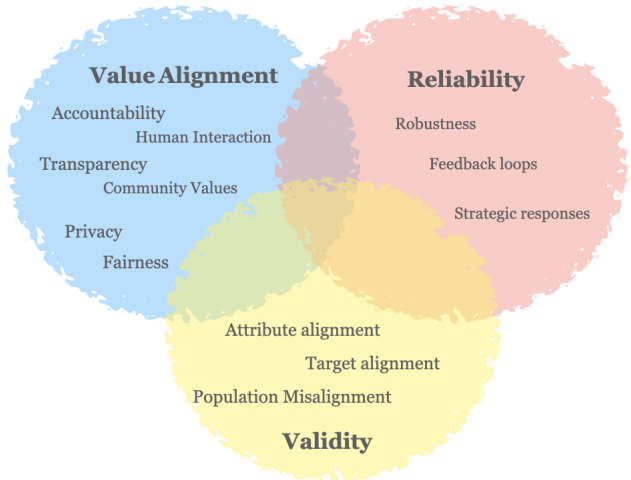

Fig. 1. The justified use of algorithms in high-stakes decision making requires at minimum that we account for validity, reliability and value alignment. These concepts are overlapping and interconnected, encompassing many aspects of responsible machine learning.

validity theory often defines construct validity as the overarching concern of validity research: construct validity integrates considerations of content, criteria, and consequences into a unified construct framework [37, 42]. Messick [37] and Gergle and Tan [38] highlight distinguishable aspects of construct validity. Below we review the definition of different aspects of construct validity, highlighting aspects that are particularly relevant in assessing the validity of data-driven decision-making algorithm.

- ***Face validity*** means that the chosen measure "appears to measure what it is supposed to measure" [38]. For example, imagine you propose to assess or predict the online satisfaction with a product on a e-commerce website by measuring the proportion of positive comments among all the purchase comments. You feel that the higher the proportion of the positive comments, the more satisfied the customers were, so "on its face" it is a valid measure or prediction target. Face validity is a very weak requirement and should be used analogously to rejecting the null in hypothesis testing: rejecting face validity allows us to conclude that the measure is not valid, but failure to reject face validity does not allow us to conclude it is valid.
- ***Convergent validity*** uses more than one measure for the same construct and then demonstrates a correlation between the two measures at the same point in time. One common way to examine convergent validity is to compare your measure with a gold-standard measure or benchmark. However, Gergle and Tan [38] warned that convergent validity can suffer from the fact that the secondary variable for comparison may have similar limitations as the measure under investigation.
- ***Discriminant validity*** tests whether measurements of two concepts that are supposed to be unrelated are, in fact, unrelated. Historically researchers have struggled to

demonstrate discriminant validity for measures of social intelligence because these measures correlate highly with measures of mental alertness [43].

- ***Predictive validity*** is a validation approach where the measure is shown to accurately predict some other conceptually related variable later in time. For example, in the context of child welfare, Vaithianathan et al. [44] demonstrated the predictive validity of Allegheny Family Screening Tool (AFST) by showing that the AFST's home removal risk score *at the time of a maltreatment referral*, was also sensitive to identifying children with a heightened risk of an emergency department (ED) visit or hospitalization because of injury *during the follow-up period*. Therefore, they argued "the risk of placement into foster care as a reasonable proxy for child harm and therefore a credible outcome for training risk stratification models for use by CPS systems" [44].

Internal validity and **external validity** are important validity considerations in experimental research [36, 38]. Internal validity is the degree to which the claims of a study hold true for the particular (often artificial) study setting, while external validity is the degree to which the claims hold true for real-world contexts, with varying cultures, different population, different technological configurations, or varying times of the day [38]. Gergle and Tan [38] discussed three common ways to bolster external validity in study design: (1) choosing a study task that is a good match for the kinds of activities in the field, (2) choosing participants for the study that are as close as possible to those in the field, and (3) assessing the similarity of the behaviors between the laboratory study and the fieldwork.

Prior work on data-driven decision-making algorithms has probed various aspects of validity threats or concerns, often using the vocabulary of "measurement error", "problem formulation", and "biases". For example, Passi and Barocas [24] chronicle how the analysts' decisions during problem formulation impacts fairness of the downstream model. Relatedly Jacobs and Wallach [26] demonstrate that how one operationalizes theoretical constructs into measurable quantities impacts fairness. Suresh and Guttag [45] also highlight measurement error in their characterization of seven types of harm in machine learning and describe other biases in representation and evaluation that can threaten validity. Representation and evaluation biases can occur when the development sample and evaluation sample, respectively, do not accurately represent who is in the target population. To the best of our knowledge, there is no prior work that proposes tools or processes centered around validity issues. Our paper aims to fill this gap by drawing on the findings in these papers to structure a validity-centered artifact intended for real-world use.

B. Value Alignment

The literature on value-alignment is vast, and we therefore focus on the works most related to our purpose of developing *artifacts*, such as documents, checklists, and software toolkits, to promote justifying the use of algorithmic systems in decision-making. Documentation artifacts designed to improve transparency and inform trust have been proposed for datasets, machine learning models, and AI products and services [46, 22, 47, 21, 48]. These artifacts document typical use cases, product/development lineage, and other important specificatons in order to promote proper use as the models, data, and services are shared and re-used across a variety of contexts. Noticing that these documentation products largely represent the perspective of algorithm developers, a recent work developed a toolkit designed to engage community advocates and activists in this process [49].

An increasingly popular mechanism is checklists for fairness and ethics in machine learning. Checklists can provide a structured form for individual advocates to raise fairness or ethics concerns, but a compliance-oriented checklist may fail to capture the nuances of complex fairness and ethical challenges [19]. Recent work has advocated for checklists designed to promote conversations about ethical challenges [50]. However, checklist-style "yes or no" questions may be ill-suited for promoting deliberation. Moreover, in centering around the question "*have we performed all the steps necessary before releasing the model?*", checklists adopt a "deploy by default" framing that may encourage practitioners to err on the side of brushing concerns aside. To address these issues, we sketch a protocol to promote deliberation centered around the question *"is an algorithmic model appropriate for use in this setting?"*.

Raji et al. [20] proposed a conceptual framework, SMACTR, for developing an internal audit for algorithmic accountability throughout the machine learning development cycle. The proposed methodology is general-purpose and comprehensive, involving other documentation and checklists discussed in this section (like model cards and datasheets), but this general-purpose methodology may be complicated, expensive and time-consuming to implement, perhaps prohibitively so for teams with limited bandwidth such as the analytics division of a public sector organization. Of note, the SMACTR methodology does not focus on issues of validity. For a given class of problems (e.g., predictive analytics for decision support) there are a set of common validity issues and questions that can be detailed and re-used across contexts. Doing so would complement the SMACTR methodology.

Based on impact assessments in other domains like construction, algorithmic impact assessments (AIAs) require algorithm developers to evaluate the impacts of the proposed algorithm on society at large and particularly on marginalized communities [51, 52, 53]. In 2019 Canada made it compulsory for a government agency using an algorithm to conduct an algorithmic impact assessment [54]. A comprehensive AIA will likely need to involve deliberation about validity issues since an invalid algorithm may very well cause adverse impacts. Related to AIA is the UK Government's Data Ethics Framework which asks practitioners to perform a self-assessment of their transparency, fairness, and accountability [18]. The framework asks the respondent to identify user needs, consider both the

benefits and unintended/negative consequences of the project, and to assess whether historical bias or selection bias may be present in the data. This framework is helpful in its breadth and specificity. However, the framework does not address core validity issues like proxy outcomes.

A number of toolkits are available to visualize the performance metrics and tradeoffs therein of algorithmic models. Visualization software has been developed to communicate tradeoffs to algorithm designers [55] and to display intersectional group disparities [56]. A number of fairness/ethics toolkits and code repositories are available to help researchers probe model disparities and explore potential mitigations [57, 58, 59].

A strain of the literature develops pedagogical processes for improving educational instruction of ethics issues in data science curriculum. Shen et al. [60] proposed a toolkit, Value Cards, to facilitate deliberation among computer science students and practitioners. The Value Cards largely focus on tradeoffs between performance metrics, stakeholder perspectives, and algorithmic impacts. Bates et al. [61] describes the experience of integrating ethics and critical data studies into a masters of data science program.

Guides for best practices in selecting a predictive algorithm for high-stakes settings have been proposed for public policy and healthcare settings [62, 63]. For instance, Kleinberg et al. [62] discuss conceptual issues such as target specification, measurement issues, omitted payoff bias, and selective labels. Our work connects these issues, among others, to established concepts of validity from the social sciences.

C. Reliability

As mentioned earlier, *"reliability is the extent to which measurements are* repeatable — *when different persons perform the measurements, on different occasions, under different conditions, with supposedly alternative instruments which measure the same thing"* [39]. Reliability encompasses reproducibility. Reliability is also defined as the *consistency* of measurement [64], and the *stability* of measurement results over a variety of conditions [65]. Reliability is necessary but not sufficient to ensure validity. That is, reliability of a measure does not imply its validity; however, a highly unreliable measure cannot be valid [65].

Drost [39] enumerates three main dimensions of reliability: equivalence (of measurements across a variety of tests), stability over time, and internal consistency (consistency over time). There are several general classes of reliability considerations:

- **Inter-rater reliability** assesses the degree of agreement between two or more raters in their appraisals. Low inter-rater reliability could be a potential concern in human-in-the-loop designs where human decision-makers receive the predictions of a ML model, and interpret them to reach the final decisions.
- **Test-retest reliability** assesses the degree to which test scores are consistent from one test administration to the next. Population shifts [66], feedback loops [67], and strategic responses [68] are among the threats to the test-retest reliability of risk assessment instruments.
- **Inter-method reliability** assesses the degree to which test scores are consistent when there is a variation in the methods or instruments used. For example, suppose two different models are independently trained to predict the risk of default by loan applicants. Inter-method reliability assesses whether these models often reach similar predictions for the same loan applicants. Another area in which inter-method reliability is applicable to ML is the extent to which an ML model can reproduce the decisions made by human decision-makers.
- **Internal consistency reliability**, assesses the consistency of results across items within a test. Models that make significantly different predictions for similar inputs may violate this notion of reliability.

Efforts in emerging areas such as MLOps focus on the development of practical tools to assess and ensure the reliability of data-driven predictive analytics [69, 70, 71]. While these efforts are still in their infancy, there is a growing body of work pointing to an urgent need for better tooling [69, 70]. For example, Veale et al. identified key challenges for public sector adoption of algorithmic fairness ideas and methods, highlighting the risks posed by changes in policy, data practices, or organizational structures [72]. Focusing on the private sector, Holstein et al. [73] identified what large companies need to improve fairness in machine learning, highlighting the need for "domain-specific frameworks that can help them navigate any associated complexities." In addition to the above changes, feedback loops and strategic responses can induce population shifts, also known as distribution shift or dataset shift [74]. The literature on data shift concerns the fast detection and characterization of distribution shifts, including distinguishing harmful shifts from inconsequential ones [75, 76]. An active area of research in machine learning aims to design learning algorithms that make accurate predictions even if decision subjects respond strategically to the trained model (see, e.g., [77, 68, 78, 79, 80]). Generalizing such settings, Perdomo et al. [81] propose a framework called *performative predictions*, which broadly studies settings in which the act of predicting influences the prediction target.

While our work focuses on validity issues, we hope that it serves as a jumping off point for future work on reliability artifacts for predictive analytics.

IV. Threats to Validity of Predictive Models

This section delves into common challenges that jeopardize validity. We organize these challenges into three groups: population misalignment, attribute misalignment, and target misalignment. We connect these groups to notions of validity from the social sciences mentioned in Section III.

A. Attribute Misalignment

To make meaningful predictions, we must have data on pertinent predictive factors, ideally ones for which we can point to evidence supporting the claim that they are relevant to

the predictive task at hand. The choice of which features to use in prediction has clear implications for internal, external, and construct validity. If there is no plausible causal path between the target and a feature such that any correlation is entirely spurious, the inclusion of the feature immediately challenges internal and external validity. Additionally, it can fail tests of face validity. A particularly pressing example of a prediction task that lacks face validity is the use of images of human faces to purportedly "predict" criminality [82], because an extensive body of research has disproved the pseudoscience of physiognomy and phrenology [83].

Note that validity does not require all predictive factors to have a *direct* causal relationship to the target variable. For instance, race is a well-established risk factor for COVID-19 related mortality, although the causal pathways through which race and COVID-19 mortality interact are not well-understood [84, 85]. One plausible pathway is that race is causally associated with access to healthcare, and access has a causal effect on health outcomes [86, 85]. Given the existence of such plausible causal connection, race is often invoked as an important risk factor to weigh in allocation of COVID-19 mitigation resources [87, 88].

B. Target Misalignment

In practice there is often considerable misalignment between what humans intended for the algorithm to predict and what the algorithm actually predicts. These issues of construct invalidity can lead to undesirable results after deploying the predictive algorithm.

In many settings, the desired prediction target is not easily observed, and so a proxy outcome is used in its place. For the pre-trial release task in the criminal justice setting, the desired prediction target may be criminal activity, but it is not possible to directly observe all criminal activity. Instead, algorithm designers have used proxy outcomes like re-arrests or re-arrests that resulted in convictions [34, 89]. The use of proxies in this setting is particularly problematic because there are documented biases in the criminal justice system, such as racial disparities in who is likely to be arrested [35]. These systematic biases mean the predictions are not predicting who may commit a crime but instead are predicting who may be arrested. In healthcare contexts, medical costs are sometimes used to proxy health outcomes. However, due to racial bias in quality of healthcare, these proxies systematically underestimate the severity of outcomes for black patients [3]. In other settings further complications arise when the objective of the decision making task is a function of multiple desired prediction targets. For instance, in the child welfare screening setting decision makers may want to reduce both the risk of immediate danger and the long-term risk of neglect. When the algorithm is constructed to only focus on one target, then we may suffer *omitted payoff bias* if the algorithm performs worse in practice on the combined objectives than anticipated from an evaluation on the singular objective [29].

Often we only observe outcomes under the decision taken–that is, we have bandit feedback [90]. Prediction tasks in such settings are counterfactual in nature, in the sense that we would like to predict the outcome under a proposed decision [91]. An algorithm trained to predict outcomes that were observed under historical decisions will not provide a reliable estimate of what will happen under the proposed decision if the decision causally affects the outcomes. For instance, in a child welfare screening task the goal is to predict risk of adverse child welfare outcomes if no further action is taken ("screened out" of investigation). Investigation can impact the risk of adverse outcomes if the welfare agency is able to identify family needs and provide appropriate services. A predictive algorithm that is trained on the observed outcomes without properly accounting for the effect of investigation on the outcome will screen out families who are likely to benefit from services [91]. When we have measured all factors jointly affecting the decision and the outcome, we can address treatment effects by training a counterfactual prediction model [91, 92]. When some confounding factors are unavailable for use at prediction time, as long as we have access to the full set of confounding factors in a batch dataset available for training, then we can properly account for any treatment effects in the bandit feedback setting [93]. In settings where we have unmeasured factors in both the training and test data, we can predict bounds on the partially identified prediction target using sensitivity models [94].

C. Population Misalignment

Even if we can justify our choice of predictive attributes and target variable, we can still have validity issues if the dataset does not represent the target population due to selection bias or other distribution shifts. This *population misalignment* poses a threat to a valid evaluation of the predictive algorithm because performance on the dataset may not accurately reflect performance on the target population. Notably, fairness properties such as disparities in performance metrics by demographic group can be markedly different on the target population. For example, Kallus and Zhou [95] demonstrated in the context of the New York City Stop, Question, and Frisk dataset that significant disparities in error rates persist in the target distribution (all NYC residents) even when there are no disparities in error rates on the data sample (stopped residents). In the consumer lending context Coston et al. [96] found that predictive disparities computed on the population of applicants whose loan was approved notably underestimated disparities on the full set of applicants. Misalignment between the model's performance during development and performance at deployment are clear threats to predictive and external validity.

Population misalignment occurs in practice often when the dataset examples are selectively sampled (i.e., not randomly sampled) from the target population. In a number of high-stakes settings, outcomes are only observed for a selectively biased sample of the population. In consumer lending, we only observe default outcomes for applicants whose loan was approved and funded [96]. In criminal justice, we only observe re-arrest outcomes for defendants who are released

[29]. In child welfare screening, we only observe removal from home for reports that are screened in to investigation [30]. A common but potentially invalid approach in such settings is to use the selectively labelled data to both train the predictive model and perform the evaluation, implicitly treating this sample as if it were a representative sample of the target when in reality it is not.

A promising strategy to address selection bias leverages unlabeled samples from the target distribution which are often already available or could be available under an improved data collection practice [97]. For instance, in consumer lending the features (the application information) are available for all applicants [96]. If we believe that we have measured all factors affecting both the selection mechanism and our outcome of interest (i.e., no unmeasured confounding[1]), methods are available to perform a counterfactual evaluation that estimate the performance on the full population (including both labelled and unlabelled cases) by taking advantage of techniques from causal inference [98, 91]. In settings where we suspect there are unmeasured confounding factors, we can still evaluate a predictive model against the current policy if we can identify an exogenous factor (i.e., an instrumental variable) that only affects the selection mechanism and not the outcome [99, 29].

Another common mechanism under which population misalignment arises is distribution shift due to domain transfer. For example, when expanding credit access to a new international market, a company may want to transfer a model of loan default built on its customer base in one country to the new country [100]. Because population demographics and other factors may differ between the two countries, the performance of the predictive model in the source country may not be a valid evaluation of the performance we would see in the new (target) country. When unlabeled data is available from the target domain, we may wish to reweigh the source data to make it "resemble" the target data. Under the assumption that there are no unmeasured confounding factors that affect both selection into the source/target domain and the likelihood of the outcome (known as *covariate shift*), we can use the likelihood ratio as weights to estimate the performance on the target population [101, 74]. We can also use the weights to reweigh the training data in order to retrain a model.

In practice and even with extreme diligence, it is generally not possible to ensure perfect population, target, and attribute alignment. For instance, nearly all prediction settings will suffer population misalignment due to temporal differences—the training data is observed in the past whereas the prediction task is in the future. A central question concerns the *degree* of this misalignment. As a first step towards characterizing this, we propose a deliberation process to identify and reflect on sources of misalignment in a given setting.

V. Deliberating over the Validity of Predictive Models

We propose a series of questions centered around validity to evaluate the justified use of algorithms in a given decision-making context. We next present the top-level questions, discussing them in the context of the child welfare and criminal justice settings. We note that the questions presented in this section are intended purely to illustrate the skeleton of an artifact that is guided by our systematization of concepts from validity theory. Outside the scope of the current contribution, future work designing specific sub-questions must solicit feedback from stakeholders and practitioners to ensure the questions are accessible, comprehensible, and useful.

A. The High-level Structure of A Validity-Centered Protocol

At a high level, our proposed artifact will consist of five parts. Part 1 prompts the description of the decision-making task and constructs of interest. Part 2, 3, and 4 consists of questions assessing construct validity, internal validity, and external validity. Last but not least, part 5 attempts to contextualize validity concerns within the broader set of considerations around the use of algorithms (e.g., efficiency). In what follows, we briefly sketch each section. For illustrative purposes, we provide hypothetical responses in the child welfare screening setting.

1. Description of the decision-making task. To center the deliberation around validity, the first set of questions require the respondent to describe the key constructs of interest, including the decision making objective(s), the criteria across which the decision is made, and other decision points surrounding this task. For example, in the child welfare screening setting, the answer may be as follows: *The hotline call worker determines whether to screen in a report for investigation based on details in the caller's allegations and administrative records for all individuals associated with the report. The report should be screened in if the call worker suspects the child is in immediate danger or at risk of harm or neglect in the future. Preceding this screening decision was the decision by an individual (e.g., neighbor, mandated reporter, other family member) to report to the child welfare hotline. If a report is screened in for investigation, the next major decision point is whether to offer services to the family. A decision to screen out is successful when the child is not at risk of harm or neglect.*

2. Questions assessing construct validity: At a high level, construct validity requires understanding the constructs involved (e.g., the ideal target label and attributes influencing it) and the particular cause and effect relationships among them. To assess construct validity, our protocol will include questions about the following types of validity:

- **Content validity** asks whether the operationalization of each construct of interest serve as a good measure of it. One major approach to assessing content validity is to ask the opinion of experts in the relevant fields.
- **Convergent validity:** To assess convergent validity, one must assess: Is there a standard/ground-truth measure for the construct of interest? If yes, how does that correlate with the new measure on the target population?
- **Discriminant validity:** To assess discriminant validity, one must evaluate the following: Can one think of a

[1]Also known as covariate shift [74]

concept that is related but theoretically different from the construct of interest? If yes, can the proposed measure distinguish between that concept and the construct of interest?

- **Predictive validity:** refers to the ability of a test to measure some event or outcome in the future. Therefore, to assess predictive validity, we need to ask: Is there high correlation between the results of the proposed measurement and a subsequent related behavior of interest?

One effective way to prompt the respondent to respond to the above questions is to consider what question(s) they would ask an oracle who could answer anything about the future. In our child welfare example, the answer here could be as follows: *We would ask whether the child will suffer harm or neglect in the next year.* Subsequent questions will refer to the outcomes identified in this question block as "oracle outcomes"–that is, the outcomes/events the respondent would like to ask an oracle to predict.

We follow the oracle question with questions about available outcomes in the data, how these available outcomes differ from the oracle outcome(s), and whether any of the previously stated goals are not addressed by the available outcome. These questions direct the respondent to consider for which segments of the population will the oracle and available outcomes be most likely to align and for which segments of the population will the available outcome likely diverge from the oracle outcome. A key question is when the available outcomes are observed. The answer to these questions may illuminate whether measurement error, bandit feedback, or other forms of missingness pertain to this outcome. An example answer in the child welfare screening context can be the following: *Available candidate outcomes in the data include re-referral to the hotline at a later point (e.g., within six months) or removal of the child from home within a timeframe (e.g., two years). Re-referral is a noisy proxy for the oracle outcome of harm/neglect because a re-referral can occur in the absence of any harm/neglect and, on the flip side, a child may be experiencing harm or neglect even when no re-referral is made. We expect on average a child that is re-referred to be more likely to experience harm/neglect than a child whose case is not re-referred. Re-referral is more likely to occur, regardless of underlying true risk of harm/neglect, for families of color and limited socioeconomic means [1, 102, 12]. Re-referral (or lack thereof) is observed for all reports, including those that are screened in and those that are screened out. By contrast, removal from home is only observed for reports that are screened in for investigation [91].*

A subset of the construct validity questions will direct the respondent to focus on issues of bandit feedback and treatment effects. These questions ask the respondent to consider how the decision relates to the outcome, including whether the outcome is observed under all decisions and whether the decision affects the outcome (and in what ways). For example, the respondent may describe the relationship between the decision and outcome in the child welfare screening setting as follows: *The decision is whether to screen in or screen out a case for a child maltreatment investigation. The outcome that is observed for all decisions is whether the child was later re-referred to the child welfare hotline. If the case is screened in, there are additional observed outcomes: Whether the allegations are substantiated upon investigation by a caseworker, whether the family is offered support in the form of public services, and whether the child is later placed out-of-home. These outcomes are observed under screen out only when a later report is screened in for investigation. The call screener's screening decision affects the outcome. For example, the decision to screen in a case may decrease the likelihood of observing adverse outcomes if the family receives public services that lead to improved parenting practices.*

3. Questions assessing internal validity: At a high level, internal validity is concerned with the existence of defensible *causal* relationship between features and the target label. To hone in on issues of internal validity, the respondent must identify available data features that one can plausibly claim are risk factors or protective factors for the ideal oracle outcome. The respondent must additionally provide evidence to support the claim that these are valid risk factors or protective factors for the oracle outcome. For instance, a respondent in the child welfare screening setting may identify the following as risk factors and protective factors in the data: *The data contains the results of any prior child welfare investigations, and we may suspect that a child in a case that was previously found to have child neglect in the past may be at risk for future neglect. The data also contains information on how often extended members of the family (such as the grandmother) interact with or care for the child, and regular supervision from a stable guardian may mitigate risk of child harm or neglect.*

4. Questions assessing external validity: External validity is concerned with the generalizablity of the model across persons, settings, and times. The question block focusing on external validity contains questions that require the respondent to describe the population for which data is available (*training population*), including provenance, the locale and time period for which data was observed, and whether any of the observations were filtered out of the dataset (e.g., due to missing data issues). The questions similarly direct the respondent to describe the population on which the predictive algorithm will be used (*target population*), including the anticipated time frame and geographies for which the predictive algorithm will be deployed. The respondent will also be asked to specify in what ways the training population differs from the target population. In our running child welfare example, the answer may be: *The training population is all reports to the state's child welfare hotline from 2015-2020 that were recorded in the state records system. No reports were knowingly filtered out of the dataset. The target population is all reports to the state's hotline at least for the next five years. The target population likely differs from the training population because of a change in mandatory reporting in mid 2019 that expanded the definition of mandated reporter to include teachers and sports coaches. As a result, the volume of calls to the hotline increased after the policy change and likely includes some*

reports that would not have been made absent the policy change.

6. Tradeoffs between validity and competing considerations: To prompt deliberation on how to weigh misalignments threatening validity against other considerations (such as efficiency or standardization), the next set of questions requires the respondent to articulate why a predictive algorithm may support decision making and to describe how they anticipate the predictive algorithm to complement the existing tools and information available. To ground this reflection in specifics, this section will ask respondents to precisely identify the expected benefits of the algorithm (e.g., improvements in efficiency or uncovering new patterns of risk). Continuing the child welfare example, the answer may be: *We intend for the predictive algorithm to summarize the information in the administrative records which the call screeners typically do not have sufficient time to fully parse. If the administrative records contain additional patterns of risk not captured in the allegations reported by the caller, then we anticipate the predictive algorithm may be able to flag reports that should be screened in but would otherwise be screened out.*

Target respondent: The respondent(s) we expect to deliberate and document answers to these questions are the individual(s) involved in the process of bringing data-driven algorithms into the decision-making process. These may include (but are not limited to) algorithm developers, data scientists and analysts, those responsible for algorithm procurement, management, frontline decision makers, and community members.

B. Protocol as a Mechanism for Transparency, Oversight, Conversation, & Translation

We next discuss how we envision a protocol reflecting the above structure, potentially in combination with questions from other existing protocols (e.g., focused around value alignment), can serve as a mechanism for transparency, oversight, conversation, and translation.

1) **Protocol as a mechanism of transparency.** A growing body of literature discusses the need to find better ways to empower impacted community members to shape algorithm design [103, 104, 105]. However, community members struggle to do this without sufficient insight into the internal deliberation processes. The protocol can help lower these barriers. For example, without the protocol, community members may be limited to assessing the face validity of models. Publicly shared responses to protocol questions may extend community members' knowledge to encompass a wider range of model validity measures that would otherwise be inaccessible or unknown to them.
2) **Protocol as a mechanism for oversight.** If the protocol is reviewed by an independent review board, deliberations around model validity in decision-making could be guided by standards that may reflect and align expectations across practitioners, policymakers, and community members. We draw an analogy to the research Institutional Review Board (IRB), which has a goal of "protecting [the rights and welfare of] research subjects" [106]. An independent review board for this protocol may serve to protect impacted community members, as opposed to 'research subjects.' However, the review process may be limited by human biases that challenge the consistency or the quality of review across different applications depending on the reviewer's unique set of biases.
3) **Protocol as a mechanism for conversation between multiple stakeholders.** If a diverse set of stakeholders are involved in deliberating and discussing the protocol questions, the protocol could help these conversations reach those who may not typically be involved in making model-level design decisions. For example, in some public sector agencies that use algorithmic decision support tools, frontline decision-makers, organizational leaders, and model analysts may develop beliefs and goals around the use of decision-making algorithms in silo [107, 108]. The process of responding to the protocol questions can introduce opportunities for structured, proactive modes of interactions across workers who might otherwise typically work in isolation. Engaging diverse perspectives in collaborative discussions surrounding the protocol may open opportunities for better understanding and mitigating inter-organizational value misalignments [109] that would otherwise get embedded and reinforced through the model itself.
4) **Protocol as a mechanism of translation to bridge academic-practitioner divide.** Recent research suggests that many of the concepts under the purview of our envisaged protocol may be less deliberately scrutinized by practitioners developing algorithms for decision-making in the real-world [24, 72]. The protocol may help bridge this divide between the research community and real-world practitioners. For example, this protocol could be a means for the research community to operationalize concerns related to model validity into practical questions that could guide internal deliberation processes in organizations considering the design or use of algorithms for decision-making.

C. Limitations

Our paper presents an initial step towards translating theoretical validity concepts into considerations for evaluating the justified use of predictive algorithms in practice. We sketch a structure for a deliberation protocol, targeted to guide multi-stakeholder conversations regarding whether or not to develop and use a predictive algorithm. Moving forward, we plan to empirically study practitioners' current practices around validity-related concerns. This research effort will help to ground the protocol, for example, by identifying question categories that may benefit the most from further scaffolding. Future work should also explore whether subcategories of real-world domains or types of predictive algorithms require additional or alternative considerations around validity.

Importantly, we emphasize that a validity-focused deliberation protocol is *not* sufficient on its own to justify the use of a predictive algorithm. Rather, we see the primary value of such a protocol as a means to structure and scaffold critical conversations among relevant decision-makers. Moreover, validity is just one component of evaluating the justified use of algorithms, alongside considerations related to reliability, value alignment, and beyond. Last but not least, organizations deploying algorithms should iteratively and constantly re-evaluate whether a predictive algorithm's use is justified, as the conditions for a given algorithm's justification may evolve with time.

The work in this paper was shaped by the authors' perspectives as machine learning, human-computer interaction, and quantitative social science researchers. Additionally, our experiences working with county and state public agencies over several years informed the work. In future work, we will incorporate perspectives from groups not represented among the authors, including impacted community members.

VI. Concluding Remarks

This paper provides a validity perspective on evaluating the justified use of data-driven decision-making algorithms. This perspective unites concepts of validity from the social sciences with data and problem formulation issues commonly encountered in machine learning and clarifies how these concepts apply to algorithmic decision making contexts. We situate the role of validity within the broader discussion of responsible use of machine learning in societally consequential domains. We illustrate how this perspective can inform and enhance future research by sketching a validity-centered artifact to promote and document deliberation on justified use.

Acknowledgment

We thank Alexandra Chouldechova and Motahhare Eslami for their insightful feedback on the project. Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the authors.

References

[1] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." *ProPublica*, 2016.

[3] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[4] D. A. Vyas, L. G. Eisenstein, and D. S. Jones, "Hidden in plain sight—reconsidering the use of race correction in clinical algorithms," pp. 874–882, 2020.

[5] M. Gilman, "Ai algorithms intended to root out welfare fraud often end up punishing the poor instead," 2020. [Online]. Available: https://theconversation.com/ai-algorithms-intended-to-root-out-welfare-fraud-often-end-up-punishing-the-poor-instead-139505

[6] R. N. Charette, "Michigan's midas unemployment system: Algorithm alchemy created lead, not gold," 2018. [Online]. Available: https://spectrum.ieee.org/michigans-midas-unemployment-system-algorithm-alchemy-that-turned-lead-into-gold

[7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Minority neighborhoods pay higher car insurance premiums than white areas with the same risk," *ProPublica*, 2017. [Online]. Available: https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk

[8] A. Fabris, A. Mishler, S. Gottardi, M. Carletti, M. Daicampi, G. A. Susto, and G. Silvello, "Algorithmic audit of italian car insurance: Evidence of unfairness in access and pricing," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 458–468.

[9] M. Tierney, "Dismantlings," in *Dismantlings*. Cornell University Press, 2019.

[10] E. P. Baumer and M. S. Silberman, "When the implication is not to design (technology)," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 2271–2274.

[11] D. Kluttz, J. A. Kroll, J. Burrell, and D. Mulligan, "Afog workshop panel 1: What a technical 'fix'for fairness can and can't accomplish," 2018.

[12] D. E. Roberts, "Digitizing the carceral state," *Harvard Law Review*, vol. 132, 2019.

[13] R. Benjamin, "Race after technology: Abolitionist tools for the new jim code," *Social Forces*, 2019.

[14] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.

[15] R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson, "Roles for computing in social change," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2020, pp. 252–260.

[16] S. Barocas, A. J. Biega, B. Fish, J. Niklas, and L. Stark, "When not to design, build, or deploy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 695–695.

[17] M. Minow, J. Zittrain, and J. Bowers, "Technical flaws of pretrial risk assessments raise grave concerns," *Berkman Klein Center, July*, vol. 17, pp. 2019–07, 2019.

[18] C. Digital and D. Office, "Data ethics framework," June 2018. [Online]. Available: https://www.gov.uk/government/publications/data-ethics-framework

[19] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-designing checklists to understand organizational challenges and opportunities around fairness in ai," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[20] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.

[21] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[22] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[23] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The fallacy of ai functionality," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 959–972.

[24] S. Passi and S. Barocas, "Problem formulation and fairness," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 39–48.

[25] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[26] A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 375–385.

[27] A. Narayanan, "How to recognize ai snake oil," *Arthur Miller Lecture on Science and Ethics*, 2019.

[28] B. Recht, "Machine learning has a validity problem." Mar 2022. [Online]. Available: http://www.argmin.net/2022/03/15/external-validity/

[29] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, vol. 133, no. 1, pp. 237–293, 2018.

[30] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 134–148.

[31] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Berkman Klein Center Research Publication*, 2020.

[32] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," in *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, 2021, pp. 5–17.

[33] I. Golbin, A. S. Rao, A. Hadjarian, and D. Krittman, "Responsible ai: A primer for the legal community," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 2121–2126.

[34] R. Fogliato, A. Xiang, Z. Lipton, D. Nagin, and A. Chouldechova, "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*, 2021, pp. 100–111.

[35] M. Alexander, "The new jim crow," *Ohio St. J. Crim. L.*, vol. 9, p. 7, 2011.

[36] D. T. Campbell, "Factors relevant to the validity of experiments in social settings." *Psychological bulletin*, vol. 54, no. 4, p. 297, 1957.

[37] S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning." *American psychologist*, vol. 50, no. 9, p. 741, 1995.

[38] D. Gergle and D. S. Tan, "Experimental research in hci," in *Ways of Knowing in HCI*. Springer, 2014, pp. 191–227.

[39] E. A. Drost, "Validity and reliability in social science research," *Education Research and perspectives*, vol. 38, no. 1, pp. 105–123, 2011.

[40] C. Sierra, N. Osman, P. Noriega, J. Sabater-Mir, and A. Perelló, "Value alignment: a formal approach," *arXiv preprint arXiv:2110.09240*, 2021.

[41] C. Rudin, C. Wang, and B. Coker, "The age of secrecy and unfairness in recidivism prediction," *Harvard Data Science Review*, vol. 2.1, 2020.

[42] C. Schotte, M. Maes, R. Cluydts, D. De Doncker, and P. Cosyns, "Construct validity of the beck depression inventory in a depressive population," *Journal of Affective Disorders*, vol. 46, no. 2, pp. 115–125, 1997.

[43] D. T. Campbell and D. W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological bulletin*, vol. 56, no. 2, p. 81, 1959.

[44] R. Vaithianathan, E. Putnam-Hornstein, A. Chouldechova, D. Benavides-Prado, and R. Berger, "Hospital injury encounters of children identified by a predictive risk model for screening child maltreatment referrals: evidence from the allegheny family screening tool," *JAMA pediatrics*, vol. 174, no. 11, pp. e202 770–e202 770, 2020.

[45] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021, pp. 1–9.

[46] S. Holland, A. Hosny, and S. Newman, "The dataset nutrition label," *Data Protection and Privacy: Data Protection and Democracy (2020)*, vol. 1, 2020.

[47] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 560–575.

[48] M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski *et al.*, "Factsheets: Increasing trust in ai services through supplier's declarations of conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6–1, 2019.

[49] P. Krafft, M. Young, M. Katell, J. E. Lee, S. Narayan, M. Epstein, D. Dailey, B. Herman, A. Tam, V. Guetler *et al.*, "An action-oriented ai policy toolkit for technology audits by community advocates and activists," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 772–781.

[50] M. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Prompting conversations about fairness in ai development with checklists," 2020.

[51] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic impact assessments: A practical framework for public agency accountability," *AI Now Institute*, pp. 1–22, 2018.

[52] H. L. Janssen, "An approach for a fundamental rights impact assessment to automated decision-making," *International Data Privacy Law*, 2020.

[53] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, "Algorithmic impact assessments and accountability: The co-construction of impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 735–746.

[54] G. of Canada, "Directive on automated decision-making," 2019. [Online]. Available: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

[55] B. Yu, Y. Yuan, L. Terveen, Z. S. Wu, J. Forlizzi, and H. Zhu, "Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives," in *Proceedings of the 2020 ACM designing interactive systems conference*, 2020, pp. 1245–1257.

[56] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau, "Fairvis: Visual analytics for discovering intersectional bias in machine learning," in *2019 IEEE Conference on Visual Analytics*

Science and Technology (VAST). IEEE, 2019, pp. 46–56.

[57] J. A. Adebayo *et al.*, "Fairml: Toolbox for diagnosing bias in predictive modeling," Ph.D. dissertation, Massachusetts Institute of Technology, 2016.

[58] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[59] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[60] H. Shen, W. H. Deng, A. Chattopadhyay, Z. S. Wu, X. Wang, and H. Zhu, "Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 850–861.

[61] J. Bates, D. Cameron, A. Checco, P. Clough, F. Hopfgartner, S. Mazumdar, L. Sbaffi, P. Stordy, and A. de la Vega de León, "Integrating fate/critical data studies into data science curricula: Where are we going and how do we get there?" in *Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3351095.3372832

[62] J. Kleinberg, J. Ludwig, and S. Mullainathan, "A guide to solving social problems with machine learning," Feb 2017. [Online]. Available: https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning

[63] S. Fazel and A. Wolf, "Selecting a risk assessment tool to use in practice: a 10-point guide," *Evidence-based mental health*, vol. 21, no. 2, pp. 41–43, 2018.

[64] K. A. Bollen, *Structural equations with latent variables*. John Wiley & Sons, 1989, vol. 210.

[65] J. C. Nunnally, *Psychometric theory 3E*. Tata McGraw-hill education, 1994.

[66] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[67] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, "Runaway feedback loops in predictive policing," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 160–171.

[68] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters, "Strategic classification," in *Innovations in Theoretical Computer Science*, ser. ITCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 111–122. [Online]. Available: https://doi.org/10.1145/2840728.2840730

[69] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," *arXiv preprint arXiv:2205.02302*, 2022.

[70] S. Shankar and A. Parameswaran, "Towards observability for machine learning pipelines," *arXiv preprint arXiv:2108.13557*, 2021.

[71] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe *et al.*, "Accelerating the machine learning lifecycle with mlflow." *IEEE Data Eng. Bull.*, vol. 41, no. 4, pp. 39–45, 2018.

[72] M. Veale, M. Van Kleek, and R. Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–14.

[73] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.

[74] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[75] S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[76] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–39, 2021.

[77] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu, "Strategic classification from revealed preferences," in *Proceedings of the 2018 ACM Conference on Economics and Computation*, 2018, pp. 55–70.

[78] C. Mendler-Dünner, J. C. Perdomo, T. Zrnic, and M. Hardt, "Stochastic optimization for performative prediction," *arXiv preprint arXiv:2006.06887*, 2020.

[79] Y. Shavit, B. Edelman, and B. Axelrod, "Causal strategic linear regression," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[80] L. Hu, N. Immorlica, and J. Wortman Vaughan, "The disparate effects of strategic manipulation," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 862–870.

[81] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7599–7609.

[82] X. Wu and X. Zhang, "Automated inference on criminality using face images," *arXiv preprint arXiv:1611.04135*, pp. 4038–4052, 2016.

[83] L. Stark and J. Hutson, "Physiognomic artificial intelligence," *Available at SSRN 3927300*, 2021.

[84] D. B. G. Tai, A. Shah, C. A. Doubeni, I. G. Sia, and M. L. Wieland, "The disproportionate impact of covid-

19 on racial and ethnic minorities in the united states," *Clinical Infectious Diseases*, vol. 2020, pp. 1–4, 06 2020.

[85] K. Mackey, C. K. Ayers, K. K. Kondo, S. Saha, S. M. Advani, S. Young, H. Spencer, M. Rusek, J. Anderson, S. Veazie *et al.*, "Racial and ethnic disparities in covid-19–related infections, hospitalizations, and deaths: a systematic review," *Annals of internal medicine*, vol. 174, no. 3, pp. 362–373, 2021.

[86] D. M. Gray, A. Anyane-Yeboa, S. Balzora, R. B. Issaka, and F. P. May, "Covid-19 and the other pandemic: populations made vulnerable by systemic inequity," *Nature Reviews Gastroenterology & Hepatology*, vol. 17, no. 9, pp. 520–522, 2020.

[87] H. Schmidt, L. O. Gostin, and M. A. Williams, "Is it lawful and ethical to prioritize racial minorities for covid-19 vaccines?" *Jama*, vol. 324, no. 20, pp. 2023–2024, 2020.

[88] E. Wrigley-Field, M. V. Kiang, A. R. Riley, M. Barbieri, Y.-H. Chen, K. A. Duchowny, E. C. Matthay, D. Van Riper, K. Jegathesan, K. Bibbins-Domingo *et al.*, "Geographically targeted covid-19 vaccination is more equitable and averts more deaths than age-based thresholds alone," *Science advances*, vol. 7, no. 40, p. eabj2099, 2021.

[89] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, and S. Venkatasubramanian, "It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks," *arXiv preprint arXiv:2106.05498*, 2021.

[90] A. Swaminathan and T. Joachims, "Batch learning from logged bandit feedback through counterfactual risk minimization," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1731–1755, 2015.

[91] A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova, "Counterfactual risk assessments, evaluation, and fairness," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 582–593.

[92] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *Advances in Neural Information Processing Systems*, 2017, pp. 1697–1708.

[93] A. Coston, E. Kennedy, and A. Chouldechova, "Counterfactual predictions under runtime confounding," in *Advances in Neural Information Processing Systems*, 2020.

[94] A. Rambachan, A. Coston, and E. H. Kennedy, "Counterfactual risk assessments under unmeasured confounding," 2022.

[95] N. Kallus and A. Zhou, "Residual unfairness in fair machine learning from prejudiced data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2439–2448.

[96] A. Coston, A. Rambachan, and A. Chouldechova, "Characterizing fairness over the set of good models under selective labels," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2144–2155.

[97] N. Goel, A. Amayuelas, A. Deshpande, and A. Sharma, "The importance of modeling data missingness in algorithmic fairness: A causal perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 7564–7573.

[98] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 1097–1104.

[99] H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan, "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*, 2017, pp. 275–284.

[100] A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty, "Fair transfer learning with missing protected attributes," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 91–98.

[101] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning under covariate shift." *Journal of Machine Learning Research*, vol. 10, no. 9, 2009.

[102] U. D. of Health and H. Services, "Child maltreatment 2017," 2017.

[103] P. Krafft, M. Young, M. Katell, J. E. Lee, S. Narayan, M. Epstein, D. Dailey, B. Herman, A. Tam, V. Guetler *et al.*, "An action-oriented ai policy toolkit for technology audits by community advocates and activists," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 772–781.

[104] H. Zhu, B. Yu, A. Halfaker, and L. Terveen, "Value-sensitive algorithm design: Method, case study, and lessons," *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–23, 2018.

[105] D. Martin Jr, V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac, "Participatory problem formulation for fairer machine learning through community based system dynamics," *arXiv preprint arXiv:2005.07572*, 2020.

[106] N. C. for the Protection of Human Subjects of Biomedical and B. Research, "The belmont report: Ethical principles and guidelines for the protection of human subjects of research," 1978.

[107] A. Kawakami, V. Sivaraman, H.-F. Cheng, L. Stapleton, Y. Cheng, D. Qing, A. Perer, Z. S. Wu, H. Zhu, and K. Holstein, "Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support," *arXiv preprint arXiv:2204.02310*, 2022.

[108] D. Saxena, K. Badillo-Urquiola, P. J. Wisniewski, and S. Guha, "A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2,

pp. 1–41, 2021.
[109] N. Holten Møller, I. Shklovski, and T. T. Hildebrandt, "Shifting concepts of value: Designing algorithmic decision-support systems for public services," in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 2020, pp. 1–12.