A Conceptual Framework for Investigating and Mitigating Machine Learning Measurement Bias (MLMB) in Psychological Assessment

Louis Tay

Sang Eun Woo

Purdue University

Louis Hickman

Wharton People Analytics

Brandon Booth

Sidney D'Mello

University of Colorado Boulder

Accepted at *Advances in Methods and Practices in Psychological Science*

Address for correspondence:
Louis Tay
Department of Psychological Sciences
Purdue University
703 Third Street
West Lafayette, IN 7907
stay@purdue.edu

Louis Tay, Sang Eun Woo, and Louis Hickman was supported by grant IIS 1921111 from the National Science Foundation and the 2020 SIOP Anti-Racism Grant. Sidney D'Mello and Brandon Booth was supported by grant IIS 1921087 and DRL 2019805 from the National Science Foundation. We thank Ruohan Wang for his helpful feedback comments on earlier versions of the manuscript.

Abstract

Given significant concerns about fairness and bias in the use of artificial intelligence (AI) and machine learning (ML) for psychological assessment, we provide a conceptual framework for investigating and mitigating machine learning measurement bias (MLMB) from a psychometric perspective. MLMB is defined as differential functioning of the trained ML model between subgroups. MLMB manifests empirically when a trained ML model produces different predicted score levels for different subgroups (e.g., race, gender) despite them having the same ground-truth levels for the underlying construct of interest (e.g., personality), and/or when the model yields differential predictive accuracies across the subgroups. Because the development of ML models involves both data and algorithms, both biased data and algorithm training bias are potential sources of MLMB. Data bias can occur in the form of nonequivalence between subgroups in the ground truth, platform-based construct, behavioral expression, and/or feature computing. Algorithm training bias can occur when algorithms are developed with nonequivalence in the relation between extracted features and ground truth (i.e., algorithm features are differentially used, weighted, or transformed between subgroups). We explain how these potential sources of bias may manifest during ML model development and share initial ideas for mitigating them, recognizing that new statistical and algorithmic procedures need to be developed. We also discuss how this framework clarifies MLMB but does not reduce the complexity of the issue.

A Conceptual Framework for Investigating and Mitigating Machine Learning Measurement Bias (MLMB) in Psychological Assessment

With the growing pervasiveness of artificial intelligence (AI) and machine learning (ML) applications throughout society, concerns have intensified over systematic inequalities and unfairness that such applications may create or perpetuate. Multiple examples of bias have emerged across fields and applications of algorithms in society. In the business space, an AI tool for resume screening developed by Amazon favored men over women (Dastin, 2018). In the healthcare space, a popular algorithm used to identify patients with healthcare needs generated differential risk scores for Black versus White patients (such that sicker Black patients are assigned risk scores equivalent to healthier White patients), which can cause disparities in healthcare access (Obermeyer et al., 2019). In the context of surveillance and security, facial recognition technology has been shown to be less accurate for darker-skinned females than lighter-skinned males (Buolamwini & Gebru, 2018; Najibi, 2020), which can contribute to greater racial inequities such as differential rates of false indictments.

In psychology and related fields, there has been a surge of research utilizing AI and ML to automate and/or guide assessments and decisions about people (i.e., ML-based psychological assessment, or *ML measurement* in short), which are also increasingly incorporated into practice (Adjerid & Kelley, 2018; Harlow & Oswald, 2016; Woo, Tay, & Proctor, 2020). Researchers have investigated the validity of social media data for inferring personality (Kosinski, Stillwell, & Graepel, 2013; Park et al., 2015) and depression (De Choudhury, Gamon, Counts, & Horvitz, 2013), video data for capturing emotion (Dhall, Ramana Murthy, Goecke, Joshi, & Gedeon, 2015), college applications for prospectively predicting graduation (Hutt, Gardner, Duckworth, & D'Mello, 2019), and narrative comments for estimating job performance ratings (Speer, 2018). In industry, organizations are increasingly adopting ML measures to assess applicants’ psychological attributes and aid (or even replace) human judgments (e.g., Campion, Campion, Campion, & Reider, 2016).

Across these domains, significant concerns have been raised about potential bias, unfairness, and discrimination, which are largely untested (Oswald, Behrend, Putka, & Sinar, 2020). Failure to adequately address measurement bias in ML measurements can have far-reaching effects as ML measurements are often used to guide (or even automate) high-stake decisions with real-life consequences. For example, using biased automated assessments of job interviews may disproportionately exclude underrepresented (gender/racial) minorities from a job or occupation (Booth et al., 2021). In other instances, ML measurements are increasingly applied to clinical and medical diagnosis (Hosseinifard, Moradi, & Rostami, 2013; Kononenko, 2013), and biased ML measurements can lead to inequalities in the provision of mental health care.

One significant barrier to systematic research on the topic of ML bias (also referred to as "algorithmic bias") is the lack of methodological guidelines for defining and empirically investigating ML bias in its full complexity. Given its prevalence in psychology, our paper focuses on issues of bias in ML measurements, where ML is used to infer (i.e., *assess* or *measure*) individuals' psychological attributes and experiences such as knowledge, skills, abilities, personality, attitudes, and emotions. As such, our goal is to provide an integrative conceptual framework for investigating and mitigating measurement bias in ML measurements. Within this framework, we define what machine learning measurement bias (MLMB) is, describe how it can manifest in ML measurements, and delineate key sources of MLMB to inspire strategies for mitigating bias. In doing so, we draw connections to the existing psychological literature on measurement bias.

Our paper is structured as follows: (a) we describe how ML models are typically constructed in psychology; (b) we describe measurement bias and, by extension, MLMB and how it empirically manifests in psychological assessment; and (c) we delineate the possible sources of MLMB (biases in data and algorithms) and offer possible directions for testing and mitigating these sources of MLMB. Researchers familiar with ML may wish to skip straight to the section, **Defining and Examining Machine Learning Measurement Bias**. Our goal is to explicate overarching conceptual issues (rather than statistical issues and procedures) and address growing concerns about ML bias and fairness by developing a framework for researchers and practitioners – in psychology and beyond – to identify and address MLMB.

Fundamentals of Traditional Machine Learning for Psychological Applications

This section introduces ML and key terms typically used for ML measurement (see Table 1 for terms and definitions). Arthur Samuel (1959) coined the term *machine learning*, which referred to computers, or machines, learning how to achieve a specific output using the input data provided. Simply put, ML entails the learning of a computational model (a computer program) from data rather than being explicitly programmed by humans. While ML is distinct from pattern recognition, it can be applied to learn patterns in the data (Anzai, 1992). There are two characteristics of ML. First, learning occurs because machines or computers can improve their predicted outputs over time as more data is provided. Second, learning in this context is a functional one, meaning that there is traditionally less concern about explaining how learning occurs–instead, increasing the accuracy of predicted outputs is evidence of learning. As such, ML is often thought of as advancing prediction rather than explanation in psychological science (Yarkoni & Westfall, 2017), though more recent advances also point to the importance of explanatory ML models (Gilpin et al., 2018).

[Insert Table 1]

There are multiple ML approaches (see Ayodele, 2010), but for our paper, we focus on what psychologists and social sciences have heavily used: supervised ML (see Types of ML in Table 1). In supervised ML, computers are provided examples of inputs and outputs, and algorithms are trained to develop models of the relationship between them, as illustrated in Figure 1. The goal is to automatically predict the outputs from the given inputs. Therefore, even simple logistic and linear regressions between predictor variables (i.e., inputs) and outcomes (i.e., outputs) can serve as useful ML algorithms (D'Mello, Southwell, & Gregg, 2020). Often, psychologists use ML to refer to computers learning the relation between inputs and outputs in a single time period rather than creating a continuous learning system that improves over time (as traditionally envisioned; Samuel, 1959). When seeking to evaluate the ML model, researchers apply the trained ML model to new, previously unseen evaluation data. The ML model predicts outputs (i.e., ML predictions) that are compared to the known outputs—the higher the convergence (e.g., Pearson's or Spearman rank-order correlation for continuous variables; accuracy, precision, recall, and/or F-1 scores for discrete variables) between ML predictions and known outputs, the greater the predictive accuracy.

[Insert Figure 1]

The application of ML in psychology can be further understood in terms of the characteristics of the data it is typically used to handle. This is also presented in Table 1 in our definition of ML.

Inputs

ML is typically used in psychological research when we use information gleaned from newer forms of data beyond surveys (e.g., text, video, voice) (Woo et al., 2020). These newer forms of data are usually included as inputs to predict outputs – e.g., using social media language (inputs) to predict personality (Park et al., 2015) or life satisfaction (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013) (outputs), or using video data (inputs) to predict personality (outputs) (Hickman, Tay, & Woo, 2019; Hickman et al., 2021). Unlike typical numeric data from quantitative surveys, which are organized into rows (values) and columns (variables), these alternate forms of data (e.g., text, video, voice) are often referred to as "unstructured" data (however, this term is a misnomer since these data sources do have structure, for example, pixels in an image). Nevertheless, such data typically require additional computer processing to transform them into inputs to train the ML model (Chen & Wojcik, 2016). For example, computer programs can process and transform natural language text into many quantitative features (e.g., word counts, phrase counts) for analysis (Kern et al., 2016). See the "Machine Learning data" portion of Table 1.

Additionally, the newer forms of data are often not designed and curated for a particular research purpose but are captured as part of a technological ecosystem. These "organic data" can nevertheless be applied to assess a psychological construct or address a research question. For example, emails, text messages, or social media activity are not designed to assess personality but can contain information about personality. In this context, ML may be viewed as a sieve designed to pick out relevant information from organic data. In contrast, "designed data" (see Roberts, 2011) such as surveys, interviews, and assessments are developed with the goal of obtaining specific types of information as part of the research design. For example, the use of Likert-type personality questions to assess personality.

Also, ML is usually employed when psychology researchers are seeking to create a model from substantially more inputs (i.e., predictors) than what is typically handled with standard statistical techniques. For example, when we use ordinary least squares regression, we need to have enough sample size (*n*) for a set of input predictors to uniquely estimate the regression parameters (*p*). In other words, the number of parameters (*p*) needs to be smaller than our sample size (*n*) (Faraway, 2014). In ML, there are often many more parameters to be estimated than the sample size (i.e., $p>>n$) (e.g., Joel et al., 2020; Sheetal, Feng, & Savani, 2020). Algorithms can handle $p>>n$ and includes methods to address overfitting when there are many predictors (Putka, Beaty, & Reeder, 2018). Many ML algorithms can also model nonlinear relationships between predictors and outcomes as well as capture multiple levels of interactivity among predictors (see D'Mello, Kappas, & Gratch, 2018 for illustration). Because of this, ML techniques can be useful with many predictors (i.e., $p>>n$ scenario), even when using traditional surveys (*vs*. text, video, voice). However, we note that the highly multidimentional nature of organic data often creates a scenario where we have many predictors. For example, analyzing social media text data in terms of counts of single words can lead to many predictors due to the wide variety of words each individual uses. Similarly, video data is inherently multimodal since it contains nonverbal, paraverbal, and verbal information.

Algorithms

ML algorithms essentially learn computational models (computer programs) from data. These models are designed for generalizability in that they aim to jointly optimize both fitting to the data (model fit) and generalizing to new data (model generalizability). The computational model can take on many forms, such as an equation, a set of rules, a table of probabilities, a decision tree, a forest of decision trees, or a neural network. For example, a simple algorithm may take the form of a linear regression equation that psychologists are familiar with:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + e_i \qquad \text{where } i = 1, \ldots, n \quad \text{(Equation 1)}$$

y_i denotes the output of interest, such as personality trait scores, β_0 denotes the regression intercept, β_p denotes the feature weights, x_p denotes the features, such as social media text features, used as input predictors, and e_i denotes the error term.

ML algorithms can go beyond the traditional linear modeling techniques (e.g., linear classification models) by modeling nonlinearity (e.g., random forest, support vector machines, and neural networks; see Glossary in Appendix for more information on ML algorithms). In other words, we can move beyond the assumptions of a linear relationship between inputs and output. For example, some ML algorithms can distinguish between clusters of data points arranged in the form of a target, with a bullseye (class A) and a surrounding ring (class B), in two-dimensional space by performing nonlinear transformations of the data (D'Mello, Kappas, & Gratch, 2018). Linear classification models would fail in this case because no line separates classes A and B. Several ML algorithms (e.g., decision trees) also inherently capture interactivity among predictors (e.g., high pitch predicts high extraversion but only when accompanied with loud speech – here, the interaction is between pitch and loudness). Although our discussions are focused on traditional ML models, we refer interested readers to the Appendix for information about Deep Neural Learning models.

Outputs

Another characteristic of ML is that researchers often use ML models to automatically predict outputs in new data. With supervised ML, researchers first provide *training data* with inputs and outputs to "train" ML algorithms. Once the trained ML models are created, they can be used to predict outputs (i.e., outcomes) from new input data. The outputs are typically psychological constructs of interest. For example, Park et al. (2015) trained an ML algorithm on the Facebook posts (i.e., social media language as inputs) of more than 60,000 users to predict their self-reported personality (i.e., personality as output). This ML model can then be used to automatically predict personality from Facebook users' language (i.e., inputs) (Park et al., 2015). These outputs used for training and evaluating the ML models are referred to as *ground truth* in ML parlance. For instance, in the case of personality, ground truth typically comes from self-reports (e.g., self-reported personality) or observer scores (e.g., observer-rated personality judged from social media posts). The ground truth can also come from other sources such as demographics (e.g., age, gender; Kosinski et al., 2013) or standardized test scores. Another term for ground truth is *labeled data*. For consistency with past work in psychology, we use the term ground truth (Tay, Woo, Hickman, & Saef, 2020).

In another example, which focuses on measuring health-related outputs, an ML algorithm may be trained using the occurrences and frequency of Google Searches of flu-related terms to predict flu trends within the United States (Ginsberg et al., 2009; Santillana, Zhang, Althouse, & Ayers, 2014). Ground truth in such a case would be localized CDC-reported flu statistics over time. Once the ML algorithm is trained, the goal is to automatically predict flu trends from recent frequencies of Google search terms related to the flu. An advantage of such an approach is that we can quickly obtain predicted flu statistics over time based on recent Google searches, much faster than a manual collection of flu statistics.

Further, an ML algorithm can be trained to forecast future trends using past longitudinal data (e.g., called forecasting, which uses time series analysis) (Jebb, Tay, Wang, & Huang, 2015). For example, ML algorithms have been applied to predict future COVID-19 trends; these algorithms rely on COVID-19 statistics collected across the world as ground-truth inputs (Wang, Zheng, Li, & Zhu, 2020). Therefore, the inputs are past COVID-19 statistics, and the outputs are near-future COVID-19 statistics. The ground truth for evaluating the accuracy of these predicted outputs would be future, retrospectively collected COVID-19 statistics. Similarly, forecasting with ML algorithms can be applied to psychological constructs.

At this juncture, it is important to note that ground truth does not necessarily represent objective truth, nor is the goal to engender philosophical discussions of what truth is. Researchers should view the term “ground truth” as data that is provided to ML models to predict. It has all the limitations of regular data, whether it be reliability or validity issues. To the extent that the ground truth is fallible, ML algorithms trained to predict such ground truth will provide fallible predictions (Tay et al., 2020). This is not to say that ML models should be dismissed because ground truth will never be perfect. Rather, the same types of critical evaluations applied to the typical statistical modeling of fallible outcomes in psychological research will need to be applied to ML models. For example, when using ordinary least squares regression to predict personality scores, we appraise how personality scores were obtained and if they were measured in a reliable and valid manner; the same concerns apply to ML models.

Summary

In short, (supervised) ML is typically applied in psychology – and beyond –when one or more of the following occur: (1) use of “unstructured” and/or organic data; (2) having many predictors resulting in $p>>n$; (3) modeling nonlinearity and interactivity; and (4) a goal of automatic prediction beyond training data (i.e., generalizability). This is summarized in Figure 1. Importantly, just as for traditional assessments, it is helpful to recognize that ML algorithms can exhibit measurement bias when trained on fallible inputs and outputs, and the algorithms may also be inadequate for capturing the relationship between inputs and outputs across subgroups, both of which can cause MLMB.

Defining and Examining Machine Learning Measurement Bias

In order to understand what MLMB is, we seek to first clarify that it is distinct from fairness. In the broad field of AI and ML, the term fairness is often used in concert with, or even interchangeably with, bias (e.g., Chouldechova & Roth, 2020). However, for clarity of communication and productive discussions moving forward, it is helpful to construe fairness as a broader concept that goes beyond measurement bias and MLMB. According to the *Standards for Educational and Psychological Testing* (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 2014) and the *Principles for the Validation and Use of Personnel Selection Procedures* (2018), adopted as a policy statement of the American Psychological Association (APA), fairness is a social concept. In this vein, there are multiple meanings of fairness: (1) equal group outcomes; (2) equitable treatment of all in the measurement procedure; (3) comparable access to constructs measured in a procedure; and (4) lack of measurement bias and predictive bias. Notably, it does not consider the first notion of fairness — equal group outcomes (wherein subgroups obtain the same average scores) — as relevant to measurement bias. ML-based personality assessment (or prediction) may be picking up on genuine differences between subgroups. For example, with regard to personality, self-reports of extraversion, agreeableness, and neuroticism are generally higher in women than men (Weisberg, Deyoung, & Hirsh, 2011). Nevertheless, relying on these ML models for assessment to select individuals, such as recruiting students for college or hiring employees for work, may lead to *adverse impact*—selecting people from one subgroup disproportionately more than another subgroup (e.g., Hutt et al., 2019). This can be regarded as unfair in terms of a lack of equal outcomes between subgroups and, therefore, should be scrutinized to determine whether the construct used for selection may perpetuate inequality. In other words, one should consider whether the construct used to select individuals is fair – and this is arguably independent of the assessment procedure or the trained ML models. This concept is distinct from measurement bias, which is concerned with whether assessment instruments, and by extension trained ML models, accurately reflect genuine subgroup differences (or similarities) or are biased because they include systematic error that magnifies or diminishes such differences (or similarities).

In the following sections, we first discuss how measurement bias is defined and investigated in the psychometrics literature. Then, we extend measurement bias to ML-based assessments, and we propose how MLMB may be defined and empirically examined (see Table 2 for a summary)[1].

Measurement Bias

Within the psychological measurement literature, the context of measurement bias is evaluating potential *bias in psychological measures* (e.g., personality measures, skill assessments, cognitive tests). Measurement bias is distinct from *socio-cognitive bias* which is found in human cognitive errors in judgments or attributions (e.g., West & Kenny, 2011) or human preference for the ingroup and prejudices against outgroups (e.g., Brewer, 1979).

Psychological measures are assumed to be imperfect operationalizations or proxies of constructs (e.g., personality, social skills, emotional intelligence) that are not directly observable. Therefore in typical measurement models, constructs (as *latent* variables) are visualized as circles, and assessment items (as *observed* variables or indicators) are visualized as boxes. Observed scores for the assessment usually rely on an aggregation of assessment items or indicators, and latent scores are typically inferred from confirmatory factor analysis (CFA) (Brown, 2006) or item response theory (IRT) (Drasgow & Hulin, 1990) models.

---

[1] Due to space limitations, the current article does not cover all the traditional psychometric issues regarding reliability, validity, fairness, and bias that should be carefully considered when using ML models for psychological assessment. We encourage readers to refer to both the *Standards* (American Psychological Association et al., 2014) and the *Principles* (2018) that present information on these topics. We also recommend reviews that specifically discuss psychometric properties of ML by Bleidorn and Hopwood (2019), Tay et al. (2020), and Woo et al. (2020).

[Insert Figure 2]

In this context, measurement bias is defined as a differential relationship between the latent score (i.e., psychological construct score) and the predicted observed score (i.e., predicted score derived from CFA or IRT), or *differential functioning* of the measurement tool, across subgroups (e.g., males vs. females) (Drasgow, 1984). In the presence of differential functioning, the measurement model can produce different predicted (i.e., observed) score levels for individuals belonging to different subgroups despite them having the same latent (i.e., true) score level. As an example of this, Figure 2 "MB: Case 1" depicts predicted observed scores higher for Subgroup 1 compared to Subgroup 2 across all levels of the latent score. This form of measurement bias (known as non-compensatory measurement bias) can lead to different predicted score distributions between two subgroups despite equivalent latent score distributions. Specifically, Subgroup 1 has a higher mean level in the predicted score than Subgroup 2, although there are no mean level differences in the latent scores[2].

Another example of differential functioning is illustrated in Figure 2 "MB: Case 2" which depicts different predicted slopes between Subgroup 1 compared to Subgroup 2, where the same latent score leads to different predicted observed scores (except for where the lines cross). This form of measurement bias (known as compensatory measurement bias) can lead to the same predicted mean level score distributions, despite the presence of measurement bias. However, the predicted score variance of Subgroup 1 is larger than Subgroup 2 despite them having the same latent score variance[3].

---

[2] Note that we have created simple illustrative examples where the subgroup latent score distributions are equivalent for didactic purposes; this is not a necessary condition for evaluating measurement bias. There are certainly multiple possibilities in how subgroup score distributions can differ. However, for space considerations, we do not go into details about those possibilities.

Machine Learning Measurement Bias (MLMB)

One definition of measurement is the numeric scaling of individuals along a theoretical continuum (Nunnally & Bernstein, 1994). Both CFA and IRT are measurement models that formally represent this in linking the observed numeric scores to the latent variable (i.e., construct) continuum. When applied to psychological assessments, ML can be regarded as a type of measurement model: It assigns ML-predicted numerical scores to individuals, aligning individuals along a construct continuum based on their input data. However, unlike CFA and IRT models, the construct continuum is proxied using the ground-truth score, which itself is an observed score. CFA and IRT models seek to optimize the measure's function relating the latent score to the observed score (shown in Figure 2). The ML models, on the other hand, seek to optimize the scoring algorithm's function by utilizing input data (e.g., text, videos) to predict the output ground-truth data. Therefore, while ML measurement models are analogous to CFA and IRT measurement models, they represent a fundamentally different type of measurement approach.

That said, in discussing MLMB, we focus on the relationships between the ground-truth score and the predicted ML score (illustrated in Figure 2). This is because we are seeking to evaluate whether the predicted ML score accurately reflects genuine subgroup differences (or similarities) in the ground truth or are biased because they magnify or diminish such differences (or similarities). MLMB occurs when there is a differential relationship between the ground-truth score (i.e., the observed score produced by the assessment – or the score ML model seeks to predict) and the ML-predicted score across subgroups. In addition, the focus of MLMB is on *traine*d ML models; in other words, ML models that have been developed for assessing individuals.

---

[3] Importantly, measurement bias is not defined as differences across groups in means and/or variances. Rather, score differences in groups are empirical manifestations of (thus can serve as evidence for) measurement bias. To this end, researchers have developed procedures for examining the effect of measurement bias on predicted mean score differences between subgroups. For instance, procedures have involved integrating subgroup line differences over a chosen density distribution to examine the effect of measurement bias on mean differences. It is akin to comparing subgroup mean differences assuming the same underlying distribution. See Nye and Drasgow (2011) and Stark, Chernyshenko, and Drasgow (2004). By extension, this can also impact predicted observed scores in their variances for the same latent distributions.

MLMB is defined as differential functioning of the trained ML model between subgroups. MLMB manifests empirically when a trained ML model produces different predicted score levels for individuals belonging to different subgroups (e.g., race, gender) despite them having the same ground-truth level for the underlying construct of interest (e.g., personality), and/or when the model yields differential predictive accuracies across subgroups.

Different Subgroup Predictions for Equivalent Ground-Truth Levels between Subgroups

The examples shown in the right half of Figure 2 parallel what was described for traditional measurement bias, except that the types of scores being modeled are different. In "MLMB: Case 1," the ML model predicts a higher score for Subgroup 1 compared to Subgroup 2 across all levels of the ground-truth score. In other words, individuals in different subgroups who have the same observed score systematically receive different ML predicted scores. Overall, MLMB leads to a higher predicted mean level for Subgroup 1 despite equivalent ground-truth score distributions. In "MLMB: Case 2," the ML model differentially predicts scores for Subgroup 1 and Subgroup 2 (except for where the lines cross). Overall, MLMB leads to a larger predicted score variance in Subgroup 1 compared to Subgroup 2 despite equivalent observed score distributions.

Differential Model Accuracy between Subgroups

Another way that MLMB can manifest is in terms of differential accuracy—or an ML model that is not equally accurate across subgroups. A visual way to think about this is that the prediction line between the ML predicted scores and the ground-truth score for Subgroup 1 does not fit as well for Subgroup 2. Consider that in Figure 2, "MLMB: Case 1" that the intercept is higher for Subgroup 1 than Subgroup 2; and "MLMB: Case 2" that both the slopes and intercepts differ between Subgroup 1 and Subgroup 2. Traditional procedures for testing measurement bias typically rely on CFA or IRT model-fit statistics to evaluate if the measurement models function similarly well for each subgroup (Cheung & Rensvold, 2002; Tay, Meade, & Cao, 2015). For example, one can examine whether model-data fit improves substantially when separate models are specified across two subgroups as compared to a common model, which would indicate that the measurement models are fundamentally different, and hence, that measurement bias exists.

Within ML, the notion of model functioning, or model-data fit, is tied to predictive accuracy: Are the ML model outputs equally accurate between subgroups (i.e., are the differences in the levels of convergence between predicted scores and ground-truth scores similar between subgroups)? As discussed previously and shown in Figure 1, ML algorithms are trained with inputs and outputs (i.e., ground truth), and ML predictive accuracy is evaluated in terms of how well predicted outputs match the ground truth. There are multiple metrics for evaluating the predictive accuracy of categorical outcomes (e.g., accuracy, precision, recall, F1, AUROC) (Kobayashi, Mol, Berkers, Kismihok, & Den Hartog, 2018) and continuous outcomes (e.g., correlation, R^2, mean squared error) (Putka et al., 2018), and any of these could be used to evaluate differential accuracy between subgroups of interest. One can use a variety of metrics to evaluate whether ML predictive accuracies differ between subgroups—which would be evidence of MLMB. This is represented in Figure 1 when the ML model is operationally evaluated between two subgroups.

Empirical Investigations of MLMB

Translating the conceptual meaning of MLMB to its empirical manifestation requires careful consideration of the level of analysis, along with a recognition of the strengths and weaknesses of various methodological approaches. With regard to "different subgroup predictions for equivalent ground-truth level," one can examine whether MLMB occurs at a specific range of ground-truth scores or on the overall distribution (i.e., means, variances, skewness, etc.). With regard to "differential model accuracy between subgroups," one can examine whether MLMB occurs at the predictive accuracy level; one can also explicitly create a (regression) model to examine whether the relationship between ground-truth scores and ML predicted scores significantly differs between subgroups. We briefly describe each of the four strategies below.

Ground-Truth Score Level

To assess MLMB, one can sample individuals from different subgroups with the same ground-truth score levels and score them using the trained ML model. If the trained ML model produces different predicted subgroup scores, MLMB exists. With this approach, it is important to assess multiple ground-truth levels because certain levels may not reveal different subgroup predicted scores (e.g., Figure 2 MLMB: Case 2 where the subgroup lines cross). A limitation of this approach is that one needs sufficient sample sizes at each ground-truth level between subgroups. Practically, this method would often entail binning individuals into different ground-truth score levels; categorizing a continuous ground-truth variable may not be statistically ideal.

Ground-Truth Distribution Level

Another way to assess MLMB is by obtaining subgroup samples with the same ground-truth distributional properties (e.g., mean and variance). Conceptually, if there is no MLMB, the trained ML model should produce equivalent predicted score distributions between subgroups (e.g., mean and variance). An advantage of this approach (as opposed to the first approach described above) is that we can infer whether the effects of MLMB translate to overall measurement bias across subgroups. For example, how might MLMB affect the mean-level inference between subgroups? At the same time, one challenge with this approach is that obtaining sufficient subgroup samples with the same ground-truth distributions can be difficult because one has to recruit matched samples (i.e., matched group designs) or create statistically matched samples from a larger sample. Another issue is that matching may reduce the representativeness of the samples[4]. Further, at this point, it is not known how well these different operational procedures can create truly equivalent ground-truth distributions.

At the ground-truth distribution level, an alternative to matching subgroup samples on their ground-truth distributions is to examine whether the ML predicted scores accurately reflect the raw unmatched observed ground-truth subgroup distributions. MLMB occurs to the extent that the ML predictions systematically enlarges or reduces the extant subgroup ground-truth distributional differences (e.g., ML predicted scores exhibit larger subgroup mean differences than ground-truth scores). An advantage of this approach is that it does not require matching subgroup samples on ground-truth distributions. A potential downside is that differences in ML predictions cannot be straightforwardly interpreted as evidence of MLMB, but needs to be compared to extant ground-truth distributional differences.

---

[4] Some readers may be concerned about potential problems of generalizability with matching samples (e.g., Thorndike, 1942). However, there are statistical methods to address matched sampling to infer causal effects (Rubin, 2006). Also, it is important to note that the goal of ML model training is not to infer causal manipulation effects on matched samples, and thus the issue of generalizability in the experimental findings is less applicable in the current context.

A general limitation of assessing MLMB at the distributional level is that the subgroup ground truth differences may be accurately recovered in terms of the ML predicted scores despite the presence of MLMB. For example, ML predicted scores show no subgroup mean level differences in Figure 2 MLMB: Case 2 due to the compensatory nature of MLMB across the ground-truth continuum. In other words, detecting subgroup distributional differences in ML predicted scores when none is expected based on equivalent subgroup ground-truth distributions would indicate MLMB; but, detecting no subgroup distributional differences in ML predicted scores are insufficient for stating that there is no MLMB.

Predictive Accuracy

In terms of model functioning, one can assess MLMB by examining whether the ML model is equally accurate across subgroups. Researchers can use one or more predictive accuracy metrics and compare subgroups. An advantage of this approach is that it does not require equivalent subgroup ground-truth distributions. That said, we need to be aware that subgroup differences in ground-truth distributions can affect the magnitude of these predictive accuracy metrics. For example, range restriction (a methodological artifact) on a subgroup's ground-truth distribution can attenuate correlation estimates (see Sackett & Yang, 2000). More research is needed to understand how different predictive accuracy estimates can be affected by different subgroup sample sizes and distributions and ways to correct them.

Modeling Ground-Truth Scores and ML Predicted Scores

Differential model functioning can also be examined by applying a (regression) model between the ground-truth scores and ML predicted scores to determine if the models between subgroups are significantly different. For example, applying a regression model to determine whether a common regression line fits both groups equally well or if there need to be two separate regression lines (as shown in Figure 2). Moderated multiple regression is typically conducted in assessing predictive bias[5] of psychological assessment (i.e., is there a differential relationship between psychological assessment scores and a criterion of interest such as cognitive ability test and academic performance; Bonaccio, Reeve, & Winford, 2012). Similar to the predictive accuracy approach, this approach does not require equivalent subgroup ground-truth distributions. However, it should also be noted that methodological artifacts (e.g., subgroup sample sizes, range restriction, subgroup means) can alter the statistical inferences (Aguinis, Culpepper, & Pierce, 2016; Aguinis & Stone-Romero, 1997).

Additional Remarks

The next section illustrates the possible sources of MLMB and corresponding mitigation strategies. The goal is to help researchers concretely understand the possible causes of MLMB and how to identify them empirically. While MLMB can empirically manifest in multiple ways, we have chosen to focus our discussion on a scenario where there is ground-truth distributional equivalence (e.g., equivalent mean levels and variances) between subgroups of interest. In the presence of equivalent ground-truth distributions across subgroups, MLMB can be manifested (thus empirically tested) in two ways. (1) Different subgroup predictions for equivalent ground-truth level: When comparing ML predicted scores between subgroups with the same ground-truth mean levels, any mean difference that emerges would indicate MLMB. In other words, MLMB is manifested in subgroup *mean-level differences*. (2) Differential model functioning between subgroups: When comparing differences in model functioning, simplifying it to the case where there is ground-truth distributional equivalence could reduce the possible methodological confounds when assessing predictive accuracy. For example, if both subgroups have similar ground-truth score variances, it is unlikely that differences in predictive accuracy are a result of a difference in their variance—where smaller variances can lead to smaller correlations due to range restriction (Sackett & Yang, 2000). In this case, MLMB is manifested in *differential predictive accuracies* between subgroups.

---

[5] In psychometric terms, predictive bias is distinct from measurement bias (Principles, 2018). Predictive bias concerns relational equivalence between predictor scores and criterion (i.e., outcomes of interest beyond the measured construct) between subgroups. For example, do tests (predictor scores) similarly predict graduate rates (criterion) between subgroups? Measurement bias concerns the relational equivalence between latent scores and predicted observed scores between subgroups. For example, does the same level of latent score on extraversion show the same predicted observed score on an extraversion assessment between subgroups? Although they are conceptually distinct, the same techniques for evaluating relational equivalence in predictive bias can be applied to MLMB.

This approach and operationalization are straightforward and illustrate the potential sources of MLMB and mitigation strategies. Notably, we do not claim that this is the only or best approach; it is simply one of many approaches, which we chose for simplicity in presentation and for reducing possible confounds.

Identifying MLMB Sources and Mitigating MLMB in Psychological Assessment

Beyond defining and operationalizing MLMB to identify its presence, it is important to understand *why* it occurs: *What are the potential sources of measurement bias in ML models*? By identifying the possible sources of MLMB, we can determine possible ways to mitigate bias. In this section, we elaborate on both the potential sources of MLMB and possible mitigation strategies. First, we expand the Brunswik (1956) lens model, typically used by psychologists to understand the workings of interpersonal perception (Hall, Pennington, & Lueders, 2013; Hinds & Joinson, 2019), and also used in social signal processing (Mehu & Scherer, 2012). Then, we contextualize the sources of MLMB–and present potential tests and mitigation strategies–through this framework.

Expanding the Brunswik Lens Model for ML Models

We use the example of ML for personality prediction as a foundation for understanding sources of MLMB in psychological assessment. A growing number of ML applications involve personality prediction (e.g., Azucar, Marengo, & Settanni, 2018; Gladstone, Matz, & Lemaire, 2019; Hickman et al., 2021; Tay et al., 2020). Importantly, this schematic is applicable to other psychological constructs of interest, including emotions (De Choudhury, Counts, & Horvitz, 2013), values (Kern, McCarthy, Chakrabarty, & Rizoiu, 2019), and well-being (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013).

[Insert Figure 3]

ML approaches have implicitly or explicitly invoked the Brunswik (1956) lens model for connecting behavioral observations to the underlying psychological construct–in this example, personality (Hall et al., 2013; Hinds & Joinson, 2019). Fundamentally, the Brunswik lens model posits that individual differences manifest in behavioral cues (e.g., written text, voice, location, nonverbals) that are used by observers to infer the latent trait. This is represented by treating ML models as "observers" of the trait via behavioral cues, as shown in Figure 3. While this representation is useful, delineating and identifying the sources of MLMB requires expanding the model by including additional aspects of supervised ML algorithms.

First, the construct measured by the input data (e.g., video data, social media data) used to train the ML model is often circumscribed and contextualized – by virtue of the method used to capture the data – compared to the broader construct of interest. This is because the platform or procedures used to measure behavior and assess the construct via ML constrain the psychological phenomenon to a specific domain or context (Tay et al., 2020). By way of analogy, self-report personality scales have been developed both to assess one's typical, in general personality and to assess one’s workplace personality (Shaffer & Postlethwaite, 2012). Similarly, using social media data to assess personality emphasizes online personality (vs. offline personality) (Marriott & Buchanan, 2014) and using video interviews in the context of a high-stakes selection is likely to engender self-presentation that colors personality (Paulhus, Westlake, Calvez, & Harms, 2013). As such, we depict the contextualized, platform-based personality construct as overlapping with but partially distinct from the broader personality construct in Figure 4.

[Insert Figure 4]

Second, the ground truth used to assess the personality construct may not match the platform-based personality construct. In our example, self-reported personality may reflect offline rather than online personality. Practically, personality measures contextualized to the situation of interest exhibit stronger relationships with situation-specific behavior than noncontextualized measures of personality (Shaffer & Postlethwaite, 2012). In other cases, other-reports of personality (e.g., roommates, colleagues, family) in everyday life may be used as ground truth, and these ratings would similarly not match the platform-based personality construct. Platform-based personality may be directly assessed (and used as ground truth instead) via domain-specific personality measures (e.g., self-report of social media personality) or observer ratings of personality based on platform activity (e.g., by trained raters of social media profiles). We note that there may be theoretical disagreement on whether there is actually a platform-based personality construct. While it could be empirically examined (e.g., assessing similarities between general self-report personality and platform-based personality) for each context, the more important point is that researchers need to be mindful of whether the construct at hand generalizes to the platform where the data are derived.

Third, the use of behavioral cues (e.g., verbal, paraverbal, and nonverbal behaviors) by ML models is mediated by features that are computed and processed from behavioral data. For example, verbal behavior on social media can be converted to features in multiple ways (Hickman et al., 2020), including by counting words in a priori dictionaries (e.g., as in Linguistic Inquiry and Word Count; J. W. Pennebaker, Boyd, Jordan, & Blackburn, 2015); counting the occurrence of words; and counting the occurrence of two- and three-word phrases (see Kern et al., 2016). In automated interviews, computers extract a discrete set of nonverbal behaviors such as facial action units (Ekman & Friesen, 1978), which can be scored based on their: activation; the intensity of their activation; and/or the co-occurrence of their activation with the activation of other facial action units (Bosch & D'Mello, 2019). Therefore, behaviors may map onto one or more features, as illustrated in Figure 4.

Finally, while ML models are generally depicted as "observers" of behavioral cues and even represented as such (e.g., Hinds & Joinson, 2019), the algorithms used to develop ML models use these behavioral cues to maximize the prediction of a fallible ground-truth measure. In other words, supervised ML algorithms do not directly recognize personality from behavioral cues. Instead, they are trained to use, weight, combine, and transform behavioral cues–as operationalized in computed features–to maximize the prediction of ground-truth scores. We depict this in Figure 4 as an arrow linking the ML model to ground-truth scores.

Identifying and Mitigating Potential Sources of MLMB

Using the expanded Brunswik lens model in Figure 4, we can now elaborate on the potential sources of MLMB. What might cause ML models to produce different subgroup score predictions despite equivalent ground-truth levels? What are the reasons ML models function differently between subgroups? These potential sources of measurement bias *may occur during ML model training* and need to be unpacked to determine what might be causing MLMB.

As presented earlier, it is helpful to recognize that ML models (as seen in Table 1) have two components: ML data and ML algorithms. Correspondingly, there are two broad potential sources of MLMB during the ML model creation: data bias and algorithm training bias. We define data bias as nonequivalence in the trait-relevant information content of ML data (i.e., ground truth, platform-based construct, behavioral expression, and feature computing) between subgroups. We define algorithm training bias as algorithms developed with nonequivalence in the relation between extracted features and ground truth (i.e., algorithm features are differentially used, weighted, or transformed between subgroups).

We should note that we view these potential sources of MLMB as underlying measurement biases themselves. In other words, biases in data and algorithm training lead to differential ML model functioning. Identifying and understanding the potential sources of MLMB–and the underlying biases–is necessary for mitigating MLMB. In Table 3 and the following, we examine potential sources of MLMB, discuss how these may manifest (i.e., leading to differences in predicted means or to differential predictive accuracies), and hence suggest ways to test for each potential source of MLMB and possible directions for mitigating these sources of bias.

Data Bias Source 1: Ground Truth in Training Data

A potential source of data bias is in the ground-truth scores used during ML model training. As shown in Figure 4, the ground truth typically uses observed data as a proxy for the latent trait. In other words, researchers use a traditional personality scale score, such as the mini-IPIP (Donnellan, Oswald, Baird, & Lucas, 2006), as ground truth to train ML models. From a psychological measurement perspective, we recognize that these measures are imperfect assessments of the latent trait[6]. Therefore, this source of bias occurs when the measure exhibits traditional measurement bias—or when it provides different scores to individuals belonging to different subgroups (e.g., gender, race) with the same latent trait scores. This potential source of bias can also occur when the measure provides different interval scores to individuals belonging to different subgroups, even when the latent trait intervals are equivalent. In other words, an increase in one unit on the scale has different meanings between subgroups. This is well-established in the measurement literature as measurement bias (Drasgow, 1984). The former type of bias (differences in predicted mean levels) is analogous to intercept differences (between the latent trait and the observed score). In contrast, the latter type of bias (differences in predictive accuracies) is analogous to slope differences (between the latent trait and the observed score, as illustrated in Figure 2; Vandenberg & Lance, 2000).

In short, this source of bias occurs when the ground-truth measure exhibits traditional measurement bias. Given the same standing on the latent personality trait, individuals may have different scores because of their subgroup membership. For example, men and women may use personality scales differently—even though men have the same latent conscientiousness score levels as women, they score higher on measured conscientiousness because the scale content asks about behaviors more frequently enacted and endorsed by men.

One way to test for ground-truth equivalence is to conduct measurement bias tests on the personality instruments used as ground truth (Tay et al., 2015; Vandenberg & Lance, 2000). When there is a measurement bias, mitigation strategies include using measurement models that are partially equivalent or excluding scale items that do not display measurement equivalence (Byrne & van de Vijver, 2010). If it is not possible to test for measurement bias, one can rely on past research and use personality scales that are known to be free from measurement bias across subgroups of interest. For example, past research has found gender measurement equivalence on the Mini-IPIP and the Big Five Inventory (Ock, McAbee, Mulfinger, & Oswald, 2020).

---

[6] This issue of imperfect measurement using human raters does not apply to situations where ML applications are used for predicting more “objective” ground-truth outputs such as CDC-reported flu statistics. However, there may be other biases in the data that are not generated by humans (e.g., due to systematic differences in data availability across subgroups and access to instruments used to generate ground-truth data between subgroups).

Beyond measurement bias in self-report instruments, bias can also occur in ground truth when one uses observer reports. This can occur either due to traditional measurement bias, as described above for self-report instruments, or due to socio-cognitive biases that occur among observers and cause them to provide scores that favor one subgroup compared to another (differences in predicted mean levels), despite the subgroups having the same trait distribution. It can also manifest in observers providing different interval scores to individuals belonging to different subgroups, even when the latent trait intervals are equivalent (i.e., differences in predictive accuracies). To mitigate possible socio-cognitive biases, standardization (e.g., rater training, ensuring the same rating procedures, using similar frames of reference) and aggregation across multiple diverse raters should be considered (e.g., Aguinis, Mazurkiewicz, & Heggestad, 2009; Bing, Whanger, Davison, & VanHook, 2004).

Another possible reason for bias is that the ML model was not trained on an equivalent ground truth distribution between subgroups. One way of thinking about this is in terms of the *representation* of subgroups on which the ML model is trained. For example, if the ML model is trained only on White individuals, likely, the predictive accuracies will not generalize to other races resulting in differences in predictive accuracies (i.e., MLMB). Importantly though, this is an empirical question that will need to be examined for each ML model because it may be the case that the ML model works equally well for different subgroups even when there is little to no representation for one or more groups in the original training data. However, if MLMB is found, one possible reason may be a lack of representation for one or more subgroups in the data used to train the ML model in the first place.

Beyond subgroup representation in the ML training sample, subgroups may not have similar ground truth distributions in the training data even when there is no measurement bias in the scale used. For example, it may be the case that MLMB occurs because even when men and women are equally represented in the sample used to train the ML model, the sample of men has a higher level of conscientiousness as compared to the sample of women. This can result in the ML model inadvertently using gender (and gender-related features) as a proxy for the latent trait and providing higher scores to men on conscientiousness as compared to women (differences in predicted mean levels) even when evaluated with subgroup representative samples (e.g., Barocas & Selbst, 2016). Similarly, this can potentially occur when, despite equivalent ground truth means, the ground truth variance is larger in men compared to women. This may result in the ML model inadvertently using more outcome-behavior relationships among men to predict conscientiousness compared to women, resulting in lower predictive accuracy among women compared to men (differences in predictive accuracies). To mitigate this, researchers can seek to use not only the same numbers of individuals for subgroups of interest (i.e., subgroup representation) but also to match subgroups of interest on ground-truth distributions during ML training[7].

---

[7] One may be concerned if it is ever possible to establish a lack of bias in ground-truth distributions. As discussed earlier, ground truth is assessed with imperfect psychological measures. Therefore, this is in part an empirical question, as we need to examine the measures used and draw on established procedures for assessing measurement bias on a case-by-case basis. Philosophically, if it is assumed that there is no possibility of establishing a lack of bias or at least the minimization of bias in the ground truth, it means that all our psychological measurements are problematic to the point that we ultimately cannot make score inferences between subgroups. This does not seem to be a tenable view of psychological measurement and psychological science.

Ground-Truth Matching to Probe Sources of MLMB. One operational approach presented here is to have equivalent ground-truth distributions (e.g., mean, variance, skewness) in the training sample to ensure that the effects of specific sources of bias are isolated. For example, under the condition that ground-truth personality scores are matched for White and Black individuals, observing behavioral differences between these groups on social media platforms can reveal a different source of bias (i.e., lack of behavioral expression equivalence). Therefore, for the purposes of testing other possible sources of bias, we present the case where subgroup samples are matched on the desired outputs (i.e., ground-truth variable) – such as personality traits – as shown in Figure 1.

From Figure 4, we note that there are cascading effects of bias in the case of ML-based psychological measurement: ground truth → platform-based construct → behavioral expression → feature computing → feature use → prediction utility. In this approach, to detect a potential source of MLMB at each stage, it is helpful to have ground-truth distributions matched and also to show that a potential source of MLMB has been resolved in the prior stages. For example, one can assess that there is no bias in the ground truth before assessing potential bias in the next stage of platform-based construct. If there is bias in the ground truth, one can seek to mitigate it and then assess potential bias in the next stage of platform-based construct. The accuracy of detecting potential bias at a specific stage would likely depend on the successful resolution of prior stages. This is summarized in Table 3 and discussed in the following.

Data Bias Source 2: Platform-based construct

Even with the same average levels of ground-truth personality (as assessed by general personality scales), platform-based personality may still have different average levels because it fundamentally depends on on how subgroups use the platform. For example, the phenomenon of

Black Twitter, where Black Americans "perform their identities through displays of cultural competence and knowledge" (p. 223) (such as using the Twitter platform as a social critique based on the shared identities of Black Americans) suggests that there is a greater level of display of Black identity on Twitter as compared to White identity (Florini, 2013, p. 223). This may result in differential levels of, say, platform-based extraversion between Black and White individuals even though ground-truth levels of self-reported extraversion (in general contexts) are similar. Further, there may be subgroup differences in the extent to which the ground-truth personality construct is related to the platform-based personality construct. For example, the use of self-reported personality scores reflective of the ground-truth construct may be differentially related to LinkedIn work personality between younger and older individuals. As compared to other major social media platforms, LinkedIn is used more by older working adults (Pew Research Center, 2018). Younger individuals may not have a career-based identity (and work-related personality) as firmly developed as compared to older working individuals (Kim & Kang, 2017). This can result in differences in the relation between the ground-truth construct (i.e., self-reported personality) and the platform-based personality construct (i.e., LinkedIn work personality) for different age groups.

At the mean level, one way to assess platform-based bias is to examine whether there are mean-level differences for subgroups on platform-based personality scores even when ground-truth scores distributions are kept equal. In other words, are there differences between subgroups on platform-based personality scores? At the predictive accuracy level, one can assess the relationship between scores on the ground-truth personality construct and the platform-based personality construct for the subgroups of interest. Basically, we are examining if there are predictive differences between subgroups on ground-truth personality scores and platform-based personality scores. A lack of difference between subgroups may be indicative of equivalence. In both cases, to test whether bias is present, it would require creating a way to assess platform-based personality, which could be achieved by having individuals self-report their platform-based personality or having observers judge individuals' personalities on the platform.

A proposed mitigation strategy is to reduce the discrepancy between the ground-truth personality construct and the platform-based personality construct. The latter can be obtained by using personality measures contextualized for the platform (e.g., social media, interview) or having observers rate personality as expressed on the platform itself (e.g., rating personality from social media posts or profiles) (Back et al., 2010). In other words, while general personality and platform-based personality constructs are distinct entities, we can ensure that ML models are trained on ground-truth scores that are closer to platform-based personality constructs. One direct possibility is to rely on only platform-based personality construct scores as the ground truth to remove this possible bias. Doing so removes the inferential gap (which contributes to possible bias) between platform-based personality (e.g., online personality) and the ground-truth personality measure. However, the same concerns in "Data Bias 1: Ground Truth" apply to platform-based personality scores (now used as ground truth). We note that this may not be a good mitigation strategy if the goal of the ML model is to predict self-report personality as assessed by general personality scales.

Data Bias Source 3: Behavioral Expression

For the same level of the platform-based personality construct, behaviors may not manifest similarly across subgroups. There may be subgroup differences in the quantity and type of behavioral expressions for the same level of a construct. For instance, women are more emotionally expressive than men (despite similar levels of reported emotions) (Kring & Gordon, 1998), and groups differ in their use of language and slang to express similar ideas (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013). Further, behaviors related to platform-based personality may differ across subgroups. For example, there may be a stronger *association* between platform-based extraversion and verbal behaviors on Twitter for Black individuals, as compared to White individuals, if they view it primarily as an outlet for identity expression (Florini, 2013).

Assuming that the ground-truth personality distribution is the same between subgroups and there is no platform-based construct bias, one way of assessing bias in behavioral expression equivalence is to examine whether there are differences in the mean levels of behavioral expressions. Because it is important not to confound computer extracted features with behavioral expressions, the focus should be on using human raters to examine base rates of behaviors (Realistically, this can only be done for a small subset of behaviors as humans may not be able to judge large numbers of behaviors reliably). For example, past reviews found that in the interview context, where machine learning is increasingly applied (Hickman et al., 2019), females smile and nod more than males (Frauendorfer & Mast, 2015). Significant differences when ground-truth distributions are matched would suggest behavioral expression bias. Bias can also be manifested at the level of predictive differences: platform-based personality scores correlate differently with behaviors between subgroups. For example, online honesty-humility levels may be differentially related to politeness behaviors on social media between Chinese as compared to Americans due to differences in how politeness is expressed (Li, Hickman, Tay, Ungar, & Guntuku, 2020). Where bias is detected, one mitigation strategy could be to exclude specific behaviors – or a set of behaviors (e.g., non-verbal behaviors) – that are shown to be biased[8].

---

[8] One may be concerned that discarding behavioral information revealing differences between subgroups can lead to less accurate measurement. Theoretically, only *biased* behavioral information should be excluded because it causes

Data Bias Source 4: Feature Computing

Importantly, expressed behaviors do not necessarily correspond one-to-one to the ML features that are processed and extracted from those behaviors. Even with equivalent behavioral expressions, processing and extracting features may not lead to equivalent levels of features between subgroups. This is because the feature computing and conversion procedures may not apply similarly across groups. For example, the use of slang and dialects may differ across groups on social media (Florini, 2013), and not all text mining processes are equally sensitive to such idiosyncrasies. Another barrier to achieving group parity in features is that subgroup differences in terms of access to reliable or quality technology (e.g., microphone, video camera, internet speed) can lead to missing data as well as differences in the features extracted (cf. Kayhan et al., 2018). For instance, in video data, individuals who have a slow internet connection (e.g., video lags) may have more unreliable feature computations encoded as compared to others who enact the same behaviors but have a fast internet connection. These same issues can also affect the relation between behavioral units and features that lead to bias across subgroups.

In some cases, the feature computation step itself relies on prior ML models, which themselves may be biased. For example, facial action units (e.g., a smile or an eye blink) are widely used as predictors in a range of ML applications (Bartlett, Littlewort, Frank, & Lee, 2014; Bosch & Dmello, 2019). Extracting these features from images entails the use of ML models, such as OpenFace (Baltru, Robinson, & Morency, 2016), which could itself be biased, thereby

---

subgroup differences when there are no actual subgroup differences; or if it exacerbates apparent differences between subgroups when there is only a small subgroup difference. For example, if the inclusion of certain nonverbal behaviors, such as having a low vocal pitch, leads to men scoring higher in dominance when there are no differences between men and women in the ground truth dominance scores, we would regard this as potentially biased behavioral information that should be excluded. That said, one should not eliminate non-biased features that reveal genuine subgroup differences.

propagating bias. Bias propagation can stack up rapidly. Consider a ML pipeline where speech is transcribed using a pretrained speech recognizer (which is a ML model) and then submitted to another pretrained ML model to compute sentiment scores (e.g., positive tone) (Zhang, Wang, & Liu, 2018), which are subsequently used as features for training an ML model to predict personality. It is widely known that automatic speech recognition algorithms are less precise for dialects relative to the majority group speech, or what is considered codified "standard" speech (Vergyri, Lamel, & Gauvain, 2010), resulting in one source of bias, which is then compounded when submitted to the sentiment ML model and, ultimately, to the personality models. Feature computing bias may also occur in the absence of ML models in the processing pipeline. For example, assume that pre-trained dictionaries, such as the Linguistic Inquiry and Word Count program (LWIC) (Pennebaker, Francis, & Booth, 2001) or the Affective Norms for English Words (ANEW) (Bradley & Lang, 1999), are used to measure sentiment in lieu of a ML model. Because these dictionary-based approaches rely on normative ratings for development or validation, any potential bias in the raters (ANEW raters were undergraduates in an Introductory Psychology course at one University so are not a representative sample) would propagate into the sentiment scores and then to the subsequent ML models that use the sentiment scores as features.

At the mean level, feature computing bias can manifest in significant feature differences even with the same level of behaviors being expressed between subgroups, assuming a lack of bias in the earlier stages (i.e., ground-truth distribution; lack of platform-based construct bias). At the predictive accuracy level, feature computing bias can occur when there are different levels of association between behaviors and computed features between subgroups of interest. One possible way to assess this is by coding key behaviors using observers (e.g., smiles, pauses in speech) and then correlating it with features that were derived from the behaviors. If differences between subgroups are found, a mitigation strategy could be to exclude features that show bias.

Algorithm Training Bias Source 1: Differential Feature Use

Even when there is no data bias, there can still be MLMB attributable to the ML algorithm training, such that different features (amongst all features included in the input data) are used to predict desired outputs across subgroups. This may arise due to human subjective decisions or errors that could introduce bias in the trained algorithms. In psychological assessment, this is akin to the choice of using different measurement models for subgroups (e.g., the scoring for subgroup A uses different scale items than subgroup B), which would be regarded as bias in psychological assessment (Vandenberg & Lance, 2000). For example, consider a simple case where different ML algorithms were developed for each subgroup; these different ML algorithms may use *different combinations of features to predict the desired outputs* across subgroups. Extending the example of Equation 1, one uses different input predictors (e.g., different social media text features) between subgroups, where Subgroup 1 may use x_1 and x_2 as predictors whereas Subgroup 2 may use x_3 and x_4.

One way of examining differential feature use is to see if a different list of features is being applied for different subgroups. For example, consider the use of two different ML models for different subgroups such that one subgroup uses only verbal features for predicting personality while another subgroup uses only facial features for predicting personality. The idea is to examine whether different subsets of features are used for each subgroup, which leads to the question of whether the algorithms trained are equivalent between subgroups. For interested readers, this is analogous to a lack of configural invariance within the measurement bias literature (Vandenberg & Lance, 2000). The mitigation strategy is to train an algorithm such that there is a common set of features being used and combined between subgroups.

Algorithm Training Bias Source 2: Differential Feature Weighting and Transformations

Even when training an ML algorithm between subgroups such that the same set of features is used and combined, bias can occur because researchers allow the trained algorithm to weight features differently between subgroups. Within psychological assessment, using different measurement models that weight scale items differently for subgroups (i.e., while using the same scale items, the scoring of subgroup A weights scale items differently from subgroup B) is another form of bias (Vandenberg & Lance, 2000). Extending the example of Equation 1, one may be using different subgroup intercepts (β_0) and/or feature weights ($\beta_1, \beta_2, \ldots, \beta_p$) such that subgroup 1 has different intercepts and/or feature weights from subgroup 2. The mitigation strategy would be to ensure that the trained algorithms are not providing different feature weights to different subgroups.

Additionally, it is also possible for researchers to apply different types of transformations to features (e.g., log transformation, normalization) that are subgroup-specific. Or, the same type of transformation but with a different mathematical function between subgroups. For example, when researchers normalize features for men and women separately before model training, if the training data for men and women have different means and variances on the features, the transformations are mathematically different between gender subgroups (Booth et al., 2021). Different subgroup transformations are regarded as a type of algorithm training bias within the MLMB framework.

When different sets of features are used, or when they are weighted or transformed differently between subgroups, this could result in differences in predicted mean levels or differences in predictive accuracy between subgroups despite them having the same ground-truth distributions. However, we note that even there are no resultant *consequential differences* (e.g., differences on predicted mean levels and/or predictive accuracies), from the perspective of psychological assessment, the trained algorithm – from a traditional psychometric perspective – is regarded as measurement biased because it is using different yardsticks (i.e., different ML models) for assessing different subgroups.

Despite this, we note that it is not uncommon to apply different ML models to different subgroups to maximize predictive accuracy for each subgroup, and there is a tension between maximizing predictive accuracy and MLMB. Another related tension is whether we are concerned primarily about the consequential differences (i.e., to what extent does the ML algorithm distort ground-truth subgroup differences in its predictions) versus the equivalence of ML models between subgroups. For example, researchers may create different ML models to mitigate the consequential differences of biased data. These key issues are elaborated on in our discussion section.

Assessing Data Biases through Differential Use and Weighting in ML Algorithms

Because the identification and mitigation approaches for data bias are likely not perfect, it is possible that biased data (e.g., biased features) are still included in the training of the ML algorithm. For example, some features may be inherently biased against women because the computed features are based on prior ML models that did not use representative data for women; these features may have “slipped through the cracks” even with the proposed mitigation strategies and were still included in ML training. Therefore, during the ML training process, one may find that certain features are not equally predictive between subgroups. For example, computed features for nonverbal behaviors may be less accurate for Black individuals than

White individuals due to biased feature computations; in turn, features for nonverbal behavior will likely be less predictive for Black individuals than White individuals.

As such, it is also possible to use ML algorithms to evaluate for specific biases in the data. One can apply the same ML algorithm for different subgroups and use a feature ablation approach (i.e., exclude specific features or sets of features in the ML algorithm) (e.g., Girshick, Donahue, Darrell, & Malik, 2014) to examine its impact on ML predicted scores. Feature ablation approaches identify potentially biased features when excluding specific features in the model reduces the discrepancy between subgroup predicted mean levels or subgroup predictive accuracies. In other words, these features may be biased because they lead to increased discrepancies between subgroups while excluding them reduces these discrepancies.

Discussion

This article provides an integrative framework for understanding what MLMB is, how it manifests, where it might originate, and what can be done to mitigate it. While we focus on delineating specific elements of MLMB investigations and associated mitigation strategies in the context of ML-based psychological assessments, many of these concepts apply to other types of ML applications. In this section, we first describe how our current framework goes beyond current bias frameworks within the computer science literature. Next, we discuss how our proposed framework brings greater clarity to the complex issues surrounding MLMB, highlighting several important tensions: maximizing predictive accuracy versus minimizing MLMB; the equivalence of ML algorithms versus consequential equivalence; and sample representativeness versus matched samples. Finally, we offer some practical guidelines for utilizing this framework to evaluate MLMB.

Comparison with Frameworks in Computer Science

Our MLMB framework has some distinctions from and similarities with current frameworks of evaluating MLMB (sometimes termed “fairness” in the computer science literature) (Gajane & Pechenizkiy, 2018). With regard to distinctions, one method for evaluating bias ("Individual Fairness"; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012) is where models are considered unbiased when they produce similar predictions for similar individuals. Notably, the similarity of individuals is based on the similarity of individuals in their input ML data regardless of subgroup membership. By contrast, the MLMB framework explicitly evaluates measurement bias based directly on known subgroup membership (e.g., age, sex, race, religion) as typically assessed in psychology. Whether ML is biased against a known subgroup is also one of the chief concerns rooted in anti-discrimination laws (e.g., Civil Rights Act of 1964; race, color, religion, sex, national origin) that is addressed through this framework.

Another approach (i.e., fairness through unawareness) emphasizes the exclusion of any protected variables (e.g., race, gender, ethnicity, nationality) within the models so that they are “blinded” to possible differences (Gajane & Pechenizkiy, 2018). This is potentially problematic given that researchers may be unaware of potential biases in all the types of data included in the first place and that input data included in the model could serve as close proxies of these protected variables (Barocas & Selbst, 2016). Instead, the MLMB framework enables researchers to explicitly examine and test for possible sources of data and algorithm training bias.

With regard to similarities, MLMB has conceptual similarity to evaluating the generalizability of ML models to different sociodemographic subgroups; the goal is to understand whether ML models applied to one sociodemographic group produce similarly accurate predictions to another group (Hutt et al., 2019). When there is no MLMB, we expect that it will be conceptually analogous to “counterfactual fairness” (Kusner, Loftus, Russell, &

SIlva, 2017), where the ML models should yield similar predicted scores when individuals are treated as their counterpart (e.g., counterfactually treating males as females should not affect the scores). The MLMB framework specifically notes that this should be the case where the same ground-truth score should yield the same predicted ML score, regardless of group membership (i.e., counterfactually treating individuals as members of other groups).

Notably, the focus of these prior mentioned frameworks emphasizes how models should be designed and tested for bias; and less on the potential sources of MLMB. Further, when some sources of MLMB are mentioned, they are not systematically developed within a holistic methodological framework (e.g., developed instead within a legal framework; Barocas & Selbst, 2016; Kleinberg, Ludwig, Mullainathan, & Sunstein, 2018). The MLMB framework goes beyond subgroup differences in outputs and predictive accuracies to explicate the possible underlying sources that can cause a lack of equivalence in ML model functioning between subgroups.

Maximizing Predictive Accuracy vs. Addressing MLMB

The proposed mitigation strategies for addressing MLMB might reduce the predictive accuracies for one subgroup in order to reach equivalence to another subgroup (Barocas & Selbst, 2016). For example, a mitigation strategy that removes features that are more predictive for one subgroup compared to another in order to achieve predictive equivalence may reduce the predictive accuracy for subgroups to the lowest common denominator (i.e., the subgroup with the lowest predictive accuracy). This creates a tension between maximizing predictive accuracy and ensuring that the ML model is equivalently predictive between subgroups (i.e., no MLMB). This is analogous to the tradeoff between equity (i.e., ML fairness; lack of MLMB) and efficiency (i.e., ML performance) discussed in economics and ethics (Le Grand, 1990).

While we do not have a solution to this tension, we believe that it is important to highlight this issue to advance future methodological and conceptual work. Methodologically, we note that the stepwise mitigation strategies are meant not only as practical procedures but also as illustrative tools to showcase how sources of MLMB can be identified and addressed. We are not beholden to these approaches and believe that future work will need to move away from stepwise strategies to simultaneous estimations that optimize multiple tasks (Dwork et al., 2012)—e.g., enhancing predictive accuracy for all subgroups while also reducing predictive accuracy differences among them. More research is now implementing *multi-objective optimization*. For example, within the personnel selection context, Pareto-optimization techniques have been developed to optimize predictor weights to maximize both performance and diversity of selected individuals (De Corte, Lievens, & Sackett, 2007; Song, Wee, & Newman, 2017). In ML, adversarial learning in neural networks strives toward a similar effect (Calmon et al., 2017). In other words, we can train ML measurement models to both maximize predictive accuracy and minimize MLMB.

Conceptually, one may ask whether there is truly a tension between maximizing predictive accuracy versus addressing MLMB. In computer science, it is not uncommon to include subgroup membership information in order to maximize predictive accuracy (e.g., Kleinberg et al., 2018). In such cases, different sets of features are (implicitly) used for each subgroup so that it is possible to achieve high predictive accuracies that are relatively similar across subgroups. Yet, this raises the question of whether the ML measurement is truly equivalent between subgroups. From a psychometric perspective, if features are viewed as assessment items, we would regard assessments that use different items for different subgroups as inherently biased because they are not assessing the same psychological construct between subgroups. If ML models are using different features to achieve similarly high levels of prediction (e.g., tone of voice for men vs. facial features for women to predict video interview performance), one can similarly regard it as a case of measurement bias because members of different groups are not being evaluated based on the same set of standards (i.e., using different yardsticks across groups).

This issue is further complicated in the ML context because even *different* computed features may represent the *same* psychological construct for different subgroups. For example, men and women tend to use different words and linguistic expressions on social media (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013); the same personality dimensions may manifest as different language usage patterns for men and women. Training separate ML models for each subgroup to pick up on these different uses of language between subgroups may lead to different ML models.

Importantly, our definition of MLMB still stands across all of the aforementioned scenarios: Differences in predicted mean scores across subgroups (given the same ground-truth mean levels) or differential predictive accuracy between subgroups can serve as empirical manifestations for MLMB. However, a lack of differential predicted mean levels or predictive accuracy is only *prima facie* evidence that the ML model is unbiased. It is possible that two entirely different models are being used to predict the same psychological attribute for different subgroups. A conceptual question is whether these computed features are similarly tapping into the same psychological construct. This requires a level of interpretability in the ML models being used and a clear conceptual rationale for what might constitute qualitatively distinct features resulting in nonequivalence (e.g., verbal behavioral features used for White people vs. nonverbal behavioral features used for Black people).

In general, this tension is related to validity concerns. As we mentioned earlier in the paper, the topic of construct validity in ML measurement goes beyond issues of predictive accuracy and MLMB. At the same time, it is also important to recognize that MLMB investigations not only provide evidence for (or against) ML measurement's generalizability (e.g., whether ML models apply equally well between subgroups) but also provide content-related validity evidence (e.g., do computed features capture the desired construct content) and predictive validity evidence (e.g., do ML scores predict important outcomes). These issues of measurement validation are relevant whenever ML models are used in a psychological context (and beyond) (Tay et al., 2020; Yarkoni & Westfall, 2017). For example, ground-truth measures that are unreliable or have questionable validity can cause problems for the ML model (Jacobucci & Grimm, 2020) and create generalizability and replicability concerns (Loken & Gelman, 2017).

Equivalence of ML Algorithms vs. Consequential Equivalence

Related to the above discussion is whether the MLMB framework should define algorithm training bias as nonequivalence of algorithms (i.e., using different features, weighting and/or transforming features differently) between subgroups. It is understandable within a traditional assessment context that the use of different measurement models (e.g., using different test items or different weighting schemes for test items) between subgroups would be viewed as measurement biased. However, ML models frequently rely on organic and naturalistic behavioral data (Xu, Zhang, & Zhou, 2020), which are arguably quite different from curated assessment items[9]. Unlike curated data from assessment items, it is challenging to fully overcome data bias with organic data, and using different ML algorithms between subgroups itself can be a way to mitigate data bias to achieve consequential equivalence (e.g., same predicted mean levels for the same ground-truth level and similar predictive accuracies).

---

[9] Unlike assessment items, these data are not easily interpretable with respect to the psychological construct (e.g., GPS coordinate data to predict personality). One cannot easily have researchers or diverse panels review assessment items to determine *a priori* (before including them in the assessment and measurement model) if there may be cultural, language, or contextual biases against specific subgroups of interest. For example, with traditional

We recognize that this argument holds merit. At the same time, including the aspect of algorithm training bias within the MLMB framework is helpful because researchers will need to seriously consider this issue and justify when they believe differences in ML algorithms are warranted to achieve consequential equivalence. For example, using qualitatively different features (i.e., verbal features for White individuals vs. facial features for Black individuals) between subgroups to predict personality may be regarded as more problematic and biased as compared to using qualitatively similar features (i.e., different language features between White and Black individuals, but these different features represent different variations of the same root word or meaning) to achieve consequential equivalence.

More importantly, algorithm training bias – and data bias– needs to be distinguished from consequential equivalence. In other words, the occurrence of MLMB is distinct from its consequences. In general, MLMB may not *always* translate into large effects in terms of consequential differences; one can find non-zero subgroup differences when examining different empirical manifestations of MLMB, but these differences may not necessarily be meaningful. For example, with very large sample sizes typically used in ML research, it is likely that one will find statistically significant differences. Still, these may not translate into large differences between subgroups on mean levels and/or predictive accuracy. Along these lines, it is also important to consider the practical effects for which the ML model is applied (e.g., are there

---

assessments, experts can review item reading levels to ensure that items are not biased against adults with low literacy (Ravens-Sieberer et al., 2014).

different selection or treatment rates between subgroups as a consequence of using ML-based predictions?). This can inform us on whether nonequivalence of ML algorithms should be treated as a significant source of bias that disadvantages a subgroup or not. Indeed, in traditional measurement bias, it is known that statistically significant differences in measurement models between subgroups do not necessarily translate into practical effects (e.g., the occurrence of measurement bias may translate into small practical effects if subgroup selection rates are similar for the same latent trait level) (Stark, Chernyshenko, & Drasgow, 2004). Moving forward, we hope that the field can recognize that MLMB (and data bias or algorithm training bias) is distinct from consequential or practical effects. Future work should seek to identify the statistical (e.g., the strength of effect size), social (e.g., the domain and context of life), and practical (i.e., type of outcome the ML predictions affects) criteria for evaluating when MLMB has significant adverse effects.

Sample Representativeness vs. Matched Samples

It is often discussed within extant MLMB frameworks that training sample representativeness is important for reducing MLMB (Barocas & Selbst, 2016; Kleinberg et al., 2018). There are several important points for clarification. First, when evaluating whether an *already trained* ML model is biased, we are less concerned about training sample representativeness (which should occur as a starting point) than distributional equivalence between subgroups in the evaluation data (e.g., same mean levels on an attribute). This is because we seek to determine whether the ML model advantages a subgroup of individuals compared to another, which is easier to examine under the condition of subgroup equivalence.

Second, when *training* an ML model, the notion of a representative sample for which one seeks to generalize can be challenging to define and practically obtain. One needs to determine the scope of representativeness (i.e., world, country, specific segment). In other words, what is the population that one seeks to justifiably represent? For example, in building an ML model to measure psychological attributes for personnel selection, it can be challenging to determine whether the population would be (a) the current demographic of workers within the organization; (b) the current demographic of the local community; (c) the current demographic of the nation; or (d) the demographic of applicants with adequate qualifications (which may not be immediately known). Moreover, there is a tension between what we term "desired representativeness" (ideal/aspirational representation of different subgroups) and "realized representativeness" (current representation of different subgroups). Representing the general population (in the current state) will tend to underrepresent minorities relative to majority group members, which may cause an ML model to disproportionately focus on the majority group's input-output relationships. For example, when Amazon trained its automated resume screening tool, it used the realized representativeness of computer programmers in its workforce (mainly men), causing biased results (Dastin, 2018). On the other hand, obtaining a representative sample with desired representativeness as a goal can also be problematic due to subjectivity it introduces (e.g., how do we decide on the ideal percentage of Asian individuals in computer science?). Conversely, our suggestion of using equivalent distributions for training ML may be more straightforward because there is perfect parity for the subgroups of interest. Arguably, this may also be a form of desired representativeness.

Third, it is unknown if sample representativeness is preferred for training ML models when seeking to address MLMB. This is especially true when there are existing inequalities or if subgroups are vastly different on the attributes of interest. For example, a concern is that differences between underrepresented minorities have lower scores than a majority group on standardized test scores may lead to increasing the chance that ML models inadvertently and indirectly use demographic information to make standardized test score predictions (i.e., introducing MLMB). In this case, representativeness could lead to biases against specific subgroups. It may be better to use samples that have subgroups with equivalent samples (i.e., equivalent trait distributions and numbers), regardless of sample representativeness (with respect to the population of interest), to train the ML model if the ultimate goal is to reduce MLMB. Notably, there may not necessarily be a tradeoff between representativeness and matched samples depending on the subpopulations of interest (i.e., there already exist similar distributions on the attribute of interest for subpopulations).

Finally, our approach for operationally matching subgroup ground-truth distributions to evaluate MLMB stems largely from a lack of ML procedures that account for different subgroup ground-truth distributions in developing predictive models. In traditional measurement bias research, matched sample distributions between subgroups are typically not required to estimate the differences in subgroup lines shown in Figure 2. For example, using CFA or IRT, simultaneous estimation can be done that accounts for different subgroup latent-score distributions; the procedure often treats the referent subgroup latent distribution as normally distributed and freely estimates the mean and variance of the other subgroup. However, given the novelty of this domain, we currently do not have similar procedures developed for investigating MLMB. Further, we believe that simplifying the case for when we have matched subgroup ground-truth distributions can help readers better understand the meaning of MLMB, the sources of MLMB, and mitigation strategies.

ML Models for Psychological Assessment vs. Other Uses

One concern that may arise is that we are advocating that ML models *throughout psychological research* should always be the same across subgroups (e.g., using and weighting features in the same manner). This is not the case. Our proposal is that in ML models explicitly used for psychological assessment, it is essential that the MLMB framework be considered and applied before being used in practice for decision-making or making claims about subgroup differences. In this context, where bias and fairness of scoring individuals and subgroups are critical, we need to evaluate the ML models for measurement bias and apply these standards to ensure that ML models function equivalently between subgroups of interest.

However, there are also many other uses of ML models beyond psychological assessment. In such contexts, applying the MLMB framework may not be relevant. For example, training different ML models for each subgroup to describe (and understand) how men and women with equivalent personality levels may express different verbal and nonverbal behaviors. It is akin to using traditional measurement models (e.g., CFA and IRT) to examine how subgroups may differ in their responses to a measurement tool. For example, such an approach has been advocated and applied to better understand cultural differences in scale responding (Cheung & Rensvold, 2000; Tay, Woo, Klafehn, & Chiu, 2010).

Practical Guidelines for Evaluating ML Models and Applications

For researchers and practitioners who use ML models, it is helpful to have practical guidance on evaluating trained ML models for MLMB based on this framework. Here, we highlight at least four ways to do this.

Foremost, because this framework examines MLMB in specific subgroups, it is important to determine whether the trained ML model has been evaluated on the subgroups of interest. For instance, while facial recognition had high levels of accuracy in general, datasets used for evaluating accuracy contained mostly lighter-skinned individuals. Aligned with this framework, Buolamwini & Gebru (2018) proposed a balanced sample of both gender and skin type to evaluate MLMB (termed 'bias' in their paper). They found that the trained ML models had high accuracies for lighter-skinned men but comparatively low accuracies for darker-skinned women. This example shows that when using a trained ML model, especially when using it to score and compare different subgroups, it is important to determine whether the ML model functions similarly across the different subgroups of interest. To evaluate for possible MLMB in trained ML models, we recommend using similar sample sizes and ground-truth characteristics for the subgroups. Researchers can evaluate this in the technical reports of the trained ML models or evaluate MLMB in specific subgroups of interest.

Second, one should consider whether the ML model was trained on data that have similar sample characteristics across subgroups. A lack of equivalence in the subgroup samples during ML model training can cause MLMB. For instance, using predominantly men samples to train Amazon's resume screener may have led to MLMB, where men were selected at a higher rate than women (Dastin, 2018). Training the ML model on a sample where men were more qualified than women can cause the ML model to inadvertently or indirectly use gender as a proxy for qualification.

Third, one should consider whether the platform data inputs (e.g., social media, internet searches) from which the ML model is trained may be used by subgroups differently, which leads to a greater likelihood of MLMB. For instance, because Black Americans use African-American Vernacular English on Twitter whereas White Americans do not, it has been found that ML models can "systematically classify content aligned with the African American English

(AAE) dialect as harmful at a higher rate than content aligned with White English (WE)" (p. 116, Ball-Burack, Lee, Cobbe, & Singh, 2021).

Fourth, as a consumer of ML-based products, it is important to ask and investigate whether the trained ML models differ across subgroups. From a psychometric perspective, this can evidence MLMB because different measures are being used to assess different subgroups of individuals. For example, subgroup norming of features in ML models (i.e., *z*-scoring ML features separately for each subgroup) is common in computer science applications. Past research using gender norming of ML features in the context of automated interviews did not find that it substantially reduced MLMB (Booth et al., 2021). More importantly, when ML models are used in the context of selection, one should recognize that subgroup norming (on race, sex, etc.) is prohibited according to the law (i.e., Civil Rights Act, 1991); subgroup norming of features in ML models likely falls within this prohibition.

Closing Remarks

In establishing this framework for MLMB, we are not proposing that all forms of bias in ML applications (e.g., whether certain psychological assessments used as ML ground truth are biased criteria or valid criteria in the first place; or the validity of using and interpreting ML models of psychological assessment) is now eminently solvable. Discussions on what criteria we are using in our ML applications and whether they are valid, biased, or unfair against specific subgroups (e.g., cognitive ability; personality) will need to occur, but they also fall beyond the scope of the current paper. More recent work has discussed validity issues when collecting, analyzing, and processing organic data (e.g., social media data, mobile sensing, etc.; Xu, Zhang, & Zhou, 2020) used in ML models but also more broadly in psychological research. Moreover, evaluating the validity of whether trained ML models accurately assess constructs and claims of what they assess (e.g., claiming ML models built on the ground truth of observer reports of personality assess personality as typically understood as self-reported personality) have also begun (Tay et al., 2020). We believe that these are issues the entire field of psychological assessment needs to grapple with and not merely ML-based psychological assessments.

We are focused narrowly on measurement bias, drawing on the psychometrics literature and its many decades of experience and established standards. It is in this realm of application that we seek to clarify what a specific form of bias (i.e., measurement bias) means, ways it may manifest, delineate the possible sources of bias, and provide initial ideas on mitigating them. In other words, the simplifying assumption is that the ground-truth construct selected is warranted – and the question lies more in whether the *assessment* of the underlying construct has measurement bias (see Table 3). Even within this specific arena, while we have sought to simplify the presentation to highlight the key ideas (e.g., using two-group comparisons; using matched ground-truth distributions), the issue of MLMB is highly complex and challenging (see Table 3). As noted, and also depicted in Figure 4, there are multiple layers to how sources of measurement bias can emerge and cascade throughout the ML measurement process. Moreover, establishing a lack of MLMB between two subgroups of interest does not necessarily generalize to other types of subgroupings. Despite these challenges, addressing bias and hence, fairness, is critical; developing a framework to investigate and address MLMB is essential to advance this cause. We also believe that providing relevant terminology and a framework that clarifies the different sources and issues in MLMB can be helpful, even if it does not reduce the complexity of addressing MLMB. And indeed, this framework serves to highlight the different tensions and issues that exist in assessing and mitigating MLMB.

The goal of this paper is to provide an integrative framework for investigating and mitigating MLMB. What do we mean by MLMB? How does it empirically manifest? Where in the process of creating ML models might MLMB be introduced? When researchers find MLMB, what can be done? The MLMB framework enables researchers to examine different sources of data and algorithm training biases and provides suggested mitigation strategies. Our goal is not to provide or identify specific statistical tests that are beyond the scope of the paper. We believe that this is only the beginning, and more work is required to further develop new methodological approaches to examine MLMB. Using this framework, one can develop new statistical and algorithmic indices and procedures. In addition, researchers will need to assess the practical effects of MLMB – beyond the statistical significance of MLMB – to understand its real-world impact. We hope that this framework will serve as foundational conceptual grounding for future work and discussions on MLMB.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899-917. doi:10.1037/amp0000190

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108*, 1045-1059.

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology, 62*, 405-438.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192-206.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Anzai, Y. (1992). *Pattern recognition and machine learning*. Academic Press, Inc.

Ayodele, T. O. (2010). Types of machine learning algorithms. In Y. Zhang (Ed.), *New advances in machine learning* (pp. 19-48). Rijeka, Croatia: InTech.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150-159. doi:10.1016/j.paid.2017.12.018

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21*(3), 372-374. doi:10.1177/0956797609360756

Ball-Burack, A., Lee, M. S. A., Cobbe, J., & Singh, J. (2021, March). Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 116-128).

Baltru, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV 2016)* (pp. 1-10). Washington, DC: IEEE.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671-732.

Bartlett, M. S., Littlewort, G. C., Frank, M. G., & Lee, K. (2014). Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology, 24*(7), 738-743.

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: a replication and extension. *Journal of Applied Psychology, 89*(1), 150-157. doi:10.1037/0021-9010.89.1.150

Bleidorn, W., & Hopwood, C. J. (2019). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review, 23*(2), 190-203. doi:10.1177/1088868318772990

Bonaccio, S., Reeve, C. L., & Winford, E. C. (2012). Text anxiety on cognitive ability test can result in differential predictive validity of academic performance. *Personality and Individual Differences, 52*(4), 497-502. doi:10.1016/j.paid.2011.11.015

Booth, B.M., Hickman L., Subburaj S.K., Tay, L., Woo, S.E, D'Mello S. (2021). Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews. In Proceedings of the 2021 International Conference on Multimodal

Interaction (ICMI ’21), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479897

Bosch, N., & Dmello, S. (2019). Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing*, 1-16. doi:10.1109/taffc.2019.2908837

Bradley, M., & Lang, P. (1999). *Affective norms for english words (anew): Stimuli, instruction manual, and affective ratings, Technical Report.* Retrieved from Gainesville, Fl:

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307-324.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence. *International Journal of Testing, 10*(2), 107-132. doi:10.1080/15305051003637306

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017, December). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3995-4004).

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*, 958–975.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21*, 458-474.

Cheung, G. W., & Rensvold, R. W. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 187-212. doi:10.1177/0022022100031002003

Cheung, G. W., & Rensvold, R. W. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the Acm, 63*, 82-89.

D’Mello, S., Kappas, A., & Gratch, J. (2018). The Affective Computing Approach to Affect Measurement. *Emotion Review, 10*(2), 174-183. doi:10.1177/1754073917696583

D’Mello, S., Southwell, R., & Gregg, J. (2020). Machine-Learned Computational Models Can Enhance the Study of Text and Discourse: A Case Study Using Eye Tracking to Model Reading Comprehension. *Discourse processes, 57*(5-6), 420-440. doi:10.1080/0163853x.2020.1739600

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

De Choudhury, M., Counts, S., & Horvitz, E. (2013). *Predicting postpartum changes in emotion and behavior via social media.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting depression via social media.* Paper presented at the International Conference on Weblogs and Social Media (ICWSM).

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *J Appl Psychol, 92*(5), 1380-1393. doi:10.1037/0021-9010.92.5.1380

Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., & Gedeon, T. (2015). *Video and Image based Emotion Recognition Challenges in the Wild*. Paper presented at the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18*, 192-203.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables. *Psychological Bulletin, 95*, 134-135. doi:10.1037/0033-2909.95.1.134

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 577-636). Palo Alto, CA: Consulting Psychologist Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through Awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214-226.

Ekman, P., & Friesen, W. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.

Faraway, J. (2014). *Linear models with R*. Boca Raton, FL: CRC Press.

Florini, S. (2013). Tweets, Tweeps, and Signifyin'. *Television & New Media, 15*(3), 223-237. doi:10.1177/1527476413480247

Frauendorfer, D., & Mast, M. S. (2015). The Impact of Nonverbal Behavior in the Job Interview. In A. Kostić & D. Chadee (Eds.), *The social psychology of nonverbal communication* (pp. 220-247). London: Palgrave Macmillan.

Gajane, P., & Pechenizkiy, M. (2018). On Formalizing Fairness in Prediction with Machine Learning. *arXiv preprint arXiv:1710.03184*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012-1014. doi:10.1038/nature07634

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data. *Psychological Science, 30*(7), 1087-1096. doi:10.1177/0956797619849435

Groves, R. (2011). “Designed Data” and “Organic Data”. https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html

Hall, J. A., Pennington, N., & Lueders, A. (2013). Impression management and formation on Facebook: A lens model approach. *New Media & Society, 16*(6), 958-982. doi:10.1177/1461444813495166

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447-457. doi:10.1037/met0000120

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, Advanced online publication*. doi:10.1037/apl0000695

Hickman, L., Tay, L., & Woo, S. E. (2019). Validity Evidence for Off-the-Shelf Language-Based Personality Assessment Using Video Interviews: Convergent and Discriminant Relationships with Self and Observer Ratings. *Personnel Assessment and Decisions, 5*(3). doi:10.25035/pad.2019.03.003

Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science, 28*, 204-211.

Hosseinifard, B., Moradi, M. H., & Rostami, R. (2013). Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer Methods and Programs in Biomedicine*, *109*(3), 339-345.

Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. (2019). Evaluating fairness and generalizability in models predicting on-time graduation from college applications. *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, 79-88.

Jacobucci, R., & Grimm, K. J. (2020). Machine Learning and Psychological Research: The Unexplored Effect of Measurement. *Perspectives on Psychologial Science, 15*(3), 809-816. doi:10.1177/1745691620902467

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: Examining and forecasting change. *Frontiers in Psychology, 6*, 727.

Joel, S., Eastwick, P. W., Allison, C. J., Arriaga, X. B., Baker, Z. G., Bar-Kalifa, E., . . . Wolf, S. (2020). Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies. *Proc Natl Acad Sci U S A, 117*(32), 19061-19071. doi:10.1073/pnas.1917036117

Kayhan, V. O., Chen, Z. C., French, K. A., Allen, T. D., Salomon, K., & Watkins, A. (2018). How honest are the signals? A protocol for validating wearable sensors. *Behavior Research Methods, 50*(1), 57-83. doi:10.3758/s13428-017-1005-4

Kern, M. L., McCarthy, P. X., Chakrabarty, D., & Rizoiu, M. A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of National Academy of Sciences, 116*(52), 26459-26464. doi:10.1073/pnas.1917942116

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21*(4), 507-525. doi:10.1037/met0000091

Kim, N., & Kang, S.-W. (2017). Older and More Engaged: The Mediating Role of Age-Linked Resources on Work Engagement. *Human Resource Management, 56*(5), 731-746. doi:10.1002/hrm.21802

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis, 10*, 113-174.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018). Text Classification for Organizational Researchers: A Tutorial. *Organizational Research Methods, 21*(3), 766-799. doi:10.1177/1094428117719322

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of National Academy of Sciences, 110*(15), 5802-5805. doi:10.1073/pnas.1218772110

Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology, 74*, 686-703.

Kusner, M., Loftus, J., Russell, C., & SIlva, R. (2017). *Counterfactual fairness.* Paper presented at the 31st Conference on Neural Information Processing Systems.

Le Grand, J. (1990). Equity versus efficiency: The elusive trade-off. *Ethics, 100*(3), 554-568.

Li, M., Hickman, L., Tay, L., Ungar, L. H., & Guntuku, S. C. (2020). Studying Politeness across cultures using English Twitter and Mandarin Weibo. In *Proceedings of the AMC on Human-Computer Interaction* (pp. 1-15).

Loken, E., & Gelman, A. (2017). Measurement error and the replicatoin crisis. *Science, 355*, 584-585.

Marriott, T. C., & Buchanan, T. (2014). The true self online: Personality correlates of preference for self-expression online, and observer ratings of personality online and offline. *Computers in Human Behavior, 32*, 171-177. doi:10.1016/j.chb.2013.11.014

Mehu, M., & Scherer, K. (2012). A psycho-ethological approach to social signal processing. *Cognitive Processing, 13*(2), 397-414.

Najibi, A. (2020). *Racial Discrimination in Face Recognition Technology*. Harvard. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The Practical Effects of Measurement Invariance: Gender Invariance in Two Big Five Personality Measures. *Assessment, 27*(4), 657-674. doi:10.1177/1073191119885018

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big Data in Industrial-Organizational Psychology and Human Resource Management: Forward Progress for Organizational Research and Practice. *Annual Review of Organizational Psychology and Organizational Behavior, 7*(1), 505-533. doi:10.1146/annurev-orgpsych-032117-104553

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., . . . Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934-952. doi:10.1037/pspp0000020

Paulhus, D. L., Westlake, B. G., Calvez, S. S., & Harms, P. D. (2013). Self-presentation style in job interviews: the role of personality and culture. *Journal of Applied Social Psychology, 43*(10), 2042-2059. doi:10.1111/jasp.12157

Pennebaker, J., Francis, M., & Booth, R. (2001). *Linguistic inquiry and word count (LIWC): A computerized text analysis program*. Mahwah NJ: Erlbaum Publishers.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Austin Texas.

Pew Research Center. (2018). *Social Media Use in 2018*. Retrieved from https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/02/PI_2018.03.01_Social-Media_FINAL.pdf

Principles for the Validation and Use of Personnel Selection Procedures. (2018). Industrial and Organizational Psychology: Perspectives on Science and Practice. *11(Supl 1)*, 2–97. doi:https://doi.org/10.1017/iop.2018.195

Putka, D. J., Beaty, A. S., & Reeder, M. C. (2018). Modern Prediction Methods: New Perspectives on a Common Problem. *Organizational Research Methods, 21*, 689-732.

Ravens-Sieberer, U., Devine, J., Bevans, K., Riley, A. W., Moon, J., Salsman, J. M., & Forrest, C. B. (2014). Subjective well-being measures for children were developed within the PROMIS project: presentation of first results. *Journal of Clinical Epidemiology*, *67*(2), 207-218.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112-118. doi:10.1037//0021-9010.85.1.112

Santillana, M., Zhang, D. W., Althouse, B. M., & Ayers, J. W. (2014). What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am J Prev Med, 47*(3), 341-347. doi:10.1016/j.amepre.2014.05.020

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. Personnel Psychology, 65(3), 445–493. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L. A., Lucas, R. E., Agrawal, M., . . . Ungar, L. H. (2013). *Characterizing geographic variation in well-being using tweets.* Paper presented at the Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), Boston, MA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L. A., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS One, 8*(9), e73791.

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*, 445-494.

Sheetal, A., Feng, Z., & Savani, K. (2020). Using Machine Learning to Generate Novel Hypotheses: Increasing Optimism About COVID-19 Makes People Less Willing to Justify Unethical Behaviors. *Psychological Science, 31*(10), 1222-1235. doi:10.1177/0956797620959594

Song, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology, 102*, 1636-1657.

Speer, A. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology, 71*, 299-333.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*, 3-46. doi:10.1177/1094428114553062

Tay, L., Woo, S. E., & D'Mello, S. (2019). Collaborative Research: AI-DCL EAGER: Understanding and Alleviating Potential Biases in Large Scale Employee Selection Systems: The Case of Automated Video Interviews. In. Purdue University & University of Colorado, Boulder: National Science Foundation.

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and Validity Issues in Machine Learning Approaches to Personality Assessment: A Focus on Social Media Text Mining. *European Journal of Personality, 34*(5), 826-844. doi:https://doi.org/10.1002/per.2290

Tay, L., Woo, S. E., Klafehn, J., & Chiu, C.-y. (2010). Conceptualizing and measuring culture: Problems and Solutions. In M. Viswanathan & G. Walford (Eds.), *The handbook of measurement: How social scientists generate, modify, and validate indicators and scales* (pp. 177-202): Sage.

Thorndike, R. L. (1942). Regression fallacies in the matched groups experiment. Psychometrika, 7, 85-102.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. doi:10.1177/109442810031002

Vergyri, D., Lamel, L., & Gauvain, J.-L. (2010). *Automatic speech recognition of multiple accented English data.* Paper presented at the Eleventh Annual Conference of the International Speech Communication Association.

Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos Solitons Fractals, 139*, 110058. doi:10.1016/j.chaos.2020.110058

Weisberg, Y. J., Deyoung, C. G., & Hirsh, J. B. (2011). Gender Differences in Personality across the Ten Aspects of the Big Five. *Frontiers in Psychology, 2*, 178. doi:10.3389/fpsyg.2011.00178

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*(2), 357-378. doi:10.1037/a0022936

Woo, S. E., Tay, L., & Proctor, R. W. (Eds.). (2020). *Big Data in Psychological Research*. Washington, D.C.: American Psychological Association.

Woo, S. E., Tay, L., Jebb, A., Ford, M., & Kern, M. (2020). Collecting big data for enhancing measurement quality. In S. Woo, L. Tay, & R. Proctor (Eds.). *Big Data in Psychological Research* (pp. 59-85). Washington D. C., APA Books.

Xu, H., Zhang, N., & Zhou, L. (2020). Validity concerns in research using organic data. *Journal of Management, 46*, 1257–1274.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychologial Science, 12*(6), 1100-1122. doi:10.1177/1745691617693393

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253.

Table 1. Machine Learning Terminologies

Terminology	Definition
Machine learning (ML)	Computers learning to achieve desired goals using data provided to them automatically. The characteristics of ML as applied in psychology typically comprise one or more of the following: (1) use of unstructured data (e.g., text, video, voice) that requires processing for use in ML; (2) many more parameters (*p*) to be estimated than sample size (*n*) ($p >> n$); (3) modeling nonlinearity and interactivity; and (4) goal of automatic prediction beyond training data (generalizability).
Machine learning (ML) data	Data provided to computers that are (automatically) processed and can be used for machine learning. We note that raw data (e.g., video files, text data) needs to be processed in order to compute the relevant features that can be used for machine learning. For example, raw data may be text data from social media posts. Preprocessing data (akin to data cleaning) will typically require identifying and converting relevant aspects to research (e.g., misspellings, acronyms, and variations for content words of interest) (see Kern et al., 2016; Hickman et al., 2020). Feature computation can then be conducted on the preprocessed data (e.g., counts of words and phrases; topics). These features may also be transformed or scaled (e.g., proportion of words) depending on the research interest. Data can broadly refer to raw data, processed data, and computed features, although computed features are directly used for machine learning rather than raw data.
Types of machine learning (ML)	The goal of the machine learning task. We list two common ML goals in psychology, although there are others. See Ayodele (2010) for more information. - Supervised learning: The computer is provided data in the form of inputs and desired outputs (i.e., ground truth). The goal is to develop a ML model that predicts the desired outputs (i.e., ground truth). One helpful analogy (and a common supervised learning algorithm) is regression: inputs are predictor variables, and the desired output is the outcome variable. - Unsupervised learning: The computer is provided inputs, and the goal is to develop a ML model to infer underlying latent variables or structures. One helpful analogy is factor analysis: inputs are the variables and we infer underlying latent factors.
Ground truth	The output in supervised ML that computers seek to predict. This can come in the form of demographic characteristics (e.g., age, gender), test scores, performance metrics, self-reported scores, other-reported scores, or (trained) observer scores.
Machine learning (ML) algorithm	Computer function used on ML data (e.g., nearest neighbor, linear regression, decision trees, support vector machines, naïve Bayes, neural

	networks) to develop an ML model. The ML algorithm is a function with parameters, such as a regression model with beta coefficients assigned to each input variable, that must be fitted to the data to optimally estimate the mapping from inputs to predicted outputs (also called “training” or “ML model training”).
Machine learning (ML) model	Machine learning is operationalized in the form of an automatic process of using ML data and a ML algorithm to produce the desired output. Therefore, the ML model comprises two components: ML data and ML algorithm. The ML model is an ML algorithm which been “trained” on the ML data, meaning that the parameters of the ML algorithm are configured such that the outputs resulting from applying the ML algorithm to the ML data optimally estimate the desired outputs.
Predictive accuracy	As applied to supervised learning, ML models produce predicted scores that are compared with ground-truth scores. The closer the predicted scores are to ground-truth scores, the higher the predictive accuracy. There are different types of predictive accuracy indices, and the type of index used depends on whether the ground-truth score is continuous (i.e., regression-type indices) or categorical (i.e., classification-type indices)
Machine learning measurement bias (MLMB)	MLMB is defined as differential functioning of the trained ML model between subgroups. MLMB can empirically manifest when a trained ML model produces different predicted score levels for individuals belonging to different subgroups (e.g., race, gender) despite them having the same ground-truth level for the underlying construct of interest (e.g., personality), and/or when the model yields differential predictive accuracies across the subgroups.
Sources of machine learning bias	Sources that can potentially contribute to MLMB. This includes different forms of data bias and algorithm training bias.
Training Data	Data with known inputs (predictors) and outputs (ground truth) used to develop (i.e., calibrate, select, and build) the ML model.
Evaluation Data	Data not used for training with known inputs (predictors) and outputs (ground truth) used to examine how accurate the trained ML model is.
Future Data	This is data with known inputs but unknown outputs where the ML model is applied to predict outputs.

Table 2. Comparison between traditional measurement bias and MLMB

Key Issues	Measurement Bias	Machine Learning Measurement Bias
Types of scores that are relevant	Predicted observed scores typically derived from confirmatory factor analysis (CFA) or item response theory (IRT) models of psychological assessments Latent scores typically derived from CFA or IRT models of psychological assessments	ML model predicted scores that are predictions produced by the ML model Ground-truth scores typically in the form of observed scores from psychological assessments
Defining bias	Defined as a differential relationship between the latent score and the predicted observed score, or differential functioning of the measurement tool, across subgroups One empirical manifestation is that the measurement model produces different scores for individuals belonging to different subgroups despite the same latent score level Another empirical manifestation is that the same measurement model does not fit subgroups equally well	Defined as differential functioning of the trained ML model between subgroups One empirical manifestation is when a trained ML model produces different predicted score levels for individuals belonging to different subgroups despite them having the same ground-truth level for the underlying construct of interest Another empirical manifestation is that the ML model yields differential predictive accuracies across the subgroups
Empirical manifestation of bias	Most typically assessed via differences in model-data fit: (a) differences in CFA fit between subgroups; (b) item-level subgroup differences in IRT fit Can also be assessed based on different model predicted scores for the same latent trait level	Ground-truth score level: Different ML predicted score levels between subgroups when subgroups have the same ground-truth score level Ground-truth distribution level: Different ML predicted score distributions (e.g., means, variances) between subgroups for equivalent subgroup ground-truth distributions. Or, the discrepancy between ML predicted subgroup score distributions and ground-truth subgroup score distributions. Predictive accuracy: Different ML model prediction accuracies (i.e., non-equivalent convergence of predicted scores and ground-truth scores) between subgroups Modeling ground-truth score and ML predicted scores: Applying (regression)

		models between ground-truth scores and ML predicted scores and finding that significantly different models are needed between subgroups

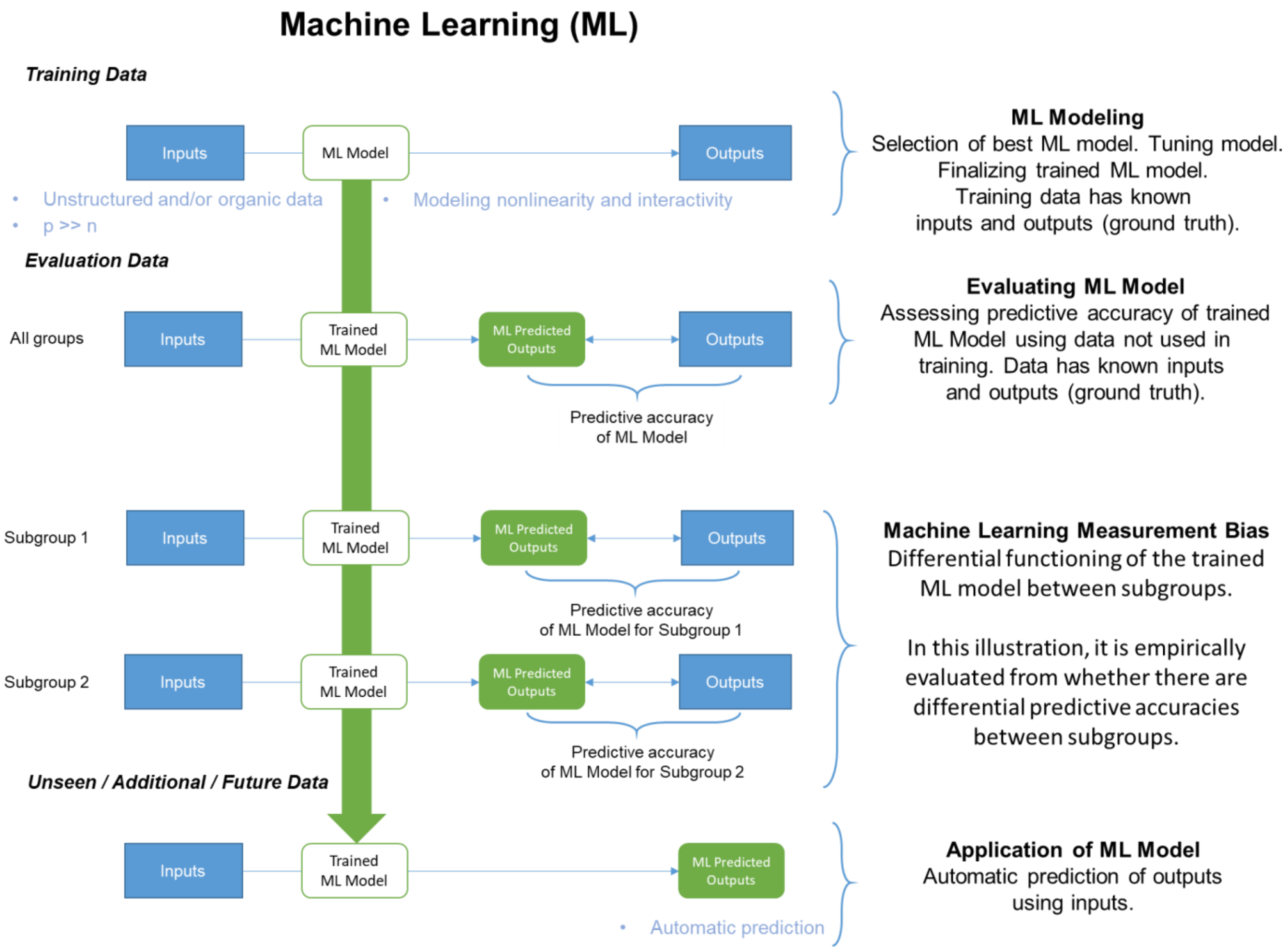
Table 3. Summary of Potential Sources of Machine Learning Bias: Identification, Testing, and Mitigation Strategy

Type of Bias	Potential Sources of Machine Learning Bias			Possible Mitigation Strategies
	Source	Illustrative underlying issues leading to different subgroup predictions for equivalent ground-truth levels, manifest as differences in predicted mean levels	Illustrative underlying issues leading to differential model functioning between subgroups, manifest as differences in predictive accuracies (e.g., differences in associations)	
Data	Ground Truth: Output in supervised ML that computers seek to predict are not equivalent between subgroups	• Measurement bias (between ground truth and latent score): Psychometric measurement bias in measures used to assess ground truth. E.g., “Men use scales differently than women that results in different subgroup scores.” • Sociocognitive bias: Sociocognitive biases in ratings and annotations. E.g., “Females are rated more severely in conscientiousness due to stereotypes.” • Sample non-equivalence in ML training (distributions or proportions): ML model trained on different ground-truth levels between subgroups may inadvertently use subgroups as a proxy in predictions. E.g., “Training sample has males with higher conscientiousness than	• Measurement bias (between ground truth and latent score): Psychometric measurement bias in measures used to assess ground truth. E.g., “Men use scales differently than women that results in different subgroup interval scores.” • Sociocognitive bias: Sociocognitive biases in ratings and annotations. E.g., “Females are have to show even more conscientiousness behaviors to be increase ratings of conscientiousness scores.” • Sample non-equivalence in ML training (distributions or proportions): ML model trained on different ground-truth variance between subgroups may inadvertently use subgroups as a proxy in predictions. E.g., “Training sample has males with higher variance in conscientiousness	• Measurement bias: One can apply partial invariant measurement models or exclude scale items that show bias. And one can seek to use measures that are known to be measurement equivalent across subgroups of interest. • Sociocognitive bias: It can be reduced through the use of training (e.g., frame of reference), standardized procedures, and the aggregation of diverse raters to assess ground truth. • Sample non-equivalence in ML training: Match subgroups of interest on ground-truth distributions in ML training. Use the same numbers of individuals for subgroups of interest in ML training.

		females”; ML model trained on considerably different proportions of subgroups may weight one subgroup more than another	than females”; ML model trained on considerably different proportions of subgroups may weight one subgroup more than another	
	Ground-Truth Matching: Matched samples on ground-truth scores are assumed for detecting the following sources of machine learning measurement biases (e.g., range restriction in a subgroup can lead to lower associations compared to another subgroup that does not have range restriction)			
	Platform-based Construct: ML model seeks to assess the same ground-truth construct between subgroups but the construct assessed in the ML model input data are not equivalent	Mean level differences on platform-based construct scores. E.g., “Black individuals show lower levels of agreeableness on social media despite having the same levels of general agreeableness as White individuals.” Assumes: - Ground-truth score distribution equivalence	Differences in the association between ground-truth construct scores and platform-based construct scores. E.g., “Black individuals have lower levels of correlation between general agreeableness and agreeableness on social media, as compared to White individuals.” Assumes: - Ground-truth score distribution equivalence	Reduce the conceptual and measurement gap between ground-truth construct and platform-based construct (e.g., use online personality rather than general personality when building ML algorithms for online personality)
	Behavioral Expression: ML model seeks to assess the same platform-based construct between subgroups but the behavioral expressions assessed in the ML model input data are not equivalent	Mean level differences on behavioral expressions. E.g., “Women display more expressions of agreeableness in an interview as compared to men, despite having the same level of agreeableness in interview settings.” Assumes: - Ground-truth score distribution equivalence	Differences in the association between platform-based construct and behavioral expressions. E.g., “Women have a smaller association between agreeableness in interview settings and expressions of agreeableness, as compared to men.” Assumes: - Ground-truth score distribution equivalence	Exclude behavioral expressions that reveal substantial differences between subgroups

		- Lack of bias in platform-based construct	- Lack of bias in platform-based construct	
	Feature Computing: ML model seeks to assess the same behavioral expressions between subgroups but the features computed in the ML model input data are not equivalent	Mean level differences on features. E.g., “Lower-income individuals have less features computed despite the same level of behavioral expressions due to poor internet connectivity.” Assumes: - Ground-truth score distribution equivalence - Lack of bias in platform-based construct - Lack of bias in behavioral expression	Differences in the association between behavioral expressions and features. E.g., “Lower-income individuals have a smaller association between behavioral expression and features computed due to poor internet connectivity.” Assumes: - Ground-truth score distribution equivalence - Lack of bias in platform-based construct - Lack of bias in behavioral expression	Exclude features that reveal substantial differences between subgroups
Algorithm Training	Feature Use, Weighting, Transformation: ML model seeks to use the same features between subgroups, but the algorithm does not treat features equivalently between subgroups	Mean level differences on ML scores when a specific feature (or feature set) is used (or differentially weighted) in the algorithm. E.g., “Using facial expression features for younger individuals and not older individuals leads to lower scores for younger individuals.” Assumes: - Ground-truth score distribution equivalence - Lack of bias in platform-based construct - Lack of bias in behavioral expression	Differences in the association between ground-truth scores and ML scores when a specific feature (or feature set) is used (or differentially weighted) in the algorithm. E.g., “Using facial expression features for younger individuals and not older individuals leads to lower prediction for younger individuals.” Assumes: - Ground-truth score distribution equivalence - Lack of bias in platform-based construct - Lack of bias in behavioral expression	Ensure a common algorithm that uses, weights, and transforms features in the same manner between subgroups Differential feature use and weighting can also be used to reveal *a priori* data bias in the features themselves. In this case, it may be appropriate to exclude features that create substantial differences in the mean level and/or predictive accuracies between subgroups (e.g., computed features for nonverbal behaviors are less accurate for Black individuals than White individuals due to non-representative prior ML models being used to compute such nonverbal features; in turn, such features may be

		- Lack of bias in feature computing	- Lack of bias in feature computing	less predictive for Black individuals than White individuals)


Machine Learning (ML)
Training Data
Inputs
ML Model
Outputs
Unstructured and/or organic data
p >> n
Modeling nonlinearity and interactivity
Evaluation Data
All groups
Inputs
Trained ML Model
ML Predicted Outputs
Outputs
Predictive accuracy of ML Model
Subgroup 1
Inputs
Trained ML Model
ML Predicted Outputs
Outputs
Predictive accuracy of ML Model for Subgroup 1
Subgroup 2
Inputs
Trained ML Model
ML Predicted Outputs
Outputs
Predictive accuracy of ML Model for Subgroup 2
Unseen / Additional / Future Data
Inputs
Trained ML Model
ML Predicted Outputs
Automatic prediction

ML Modeling
Selection of best ML model. Tuning model. Finalizing trained ML model. Training data has known inputs and outputs (ground truth).

Evaluating ML Model
Assessing predictive accuracy of trained ML Model using data not used in training. Data has known inputs and outputs (ground truth).

Machine Learning Measurement Bias
Differential functioning of the trained ML model between subgroups.

In this illustration, it is empirically evaluated from whether there are differential predictive accuracies between subgroups.

Application of ML Model
Automatic prediction of outputs using inputs.

Figure 1. Simplified process of machine learning modeling

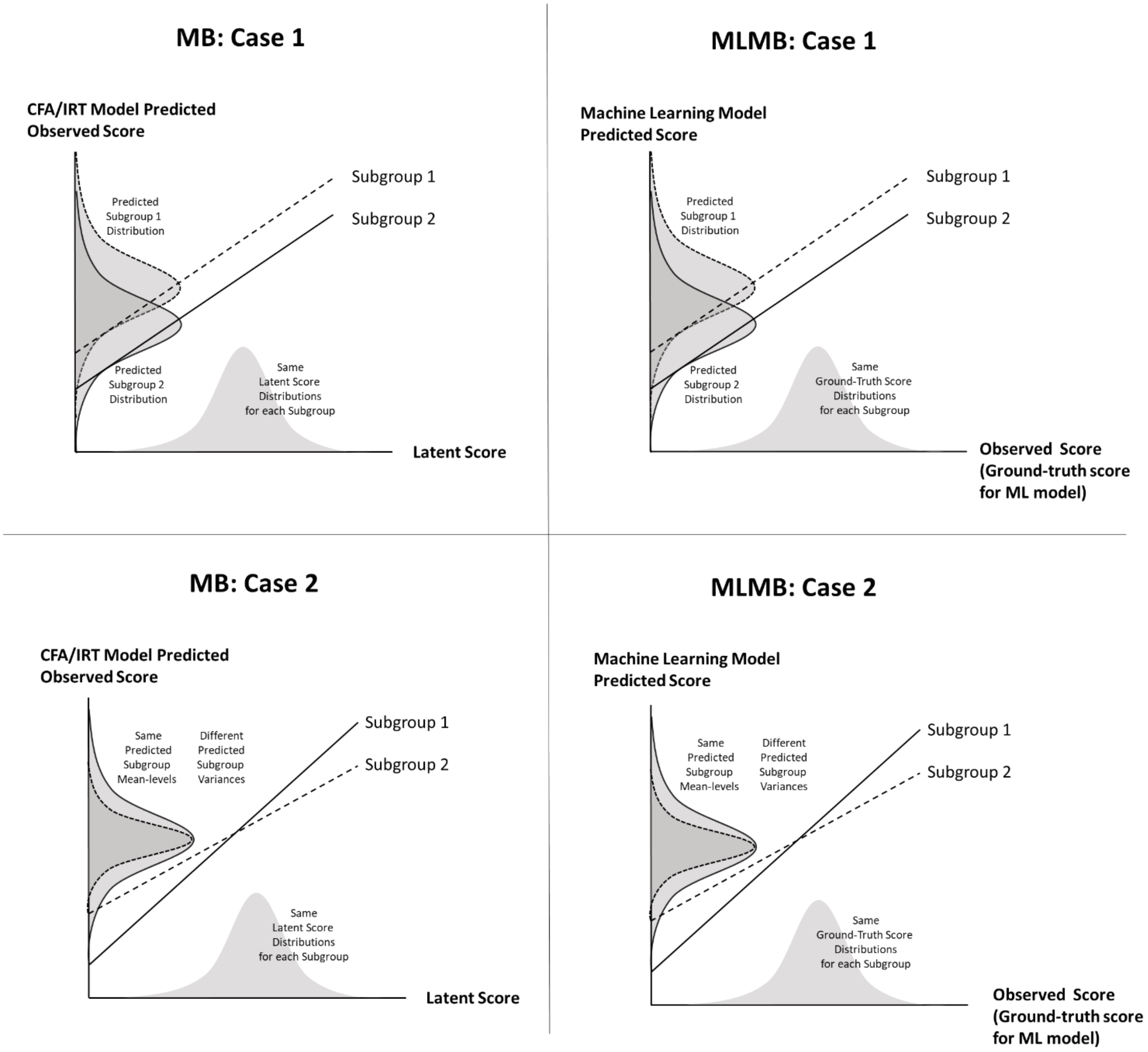

MB: Case 1
CFA/IRT Model Predicted Observed Score
Predicted Subgroup 1 Distribution
Subgroup 1
Subgroup 2
Predicted Subgroup 2 Distribution
Same Latent Score Distributions for each Subgroup
Latent Score
MLMB: Case 1
Machine Learning Model Predicted Score
Predicted Subgroup 1 Distribution
Subgroup 1
Subgroup 2
Predicted Subgroup 2 Distribution
Same Ground-Truth Score Distributions for each Subgroup
Observed Score (Ground-truth score for ML model)
MB: Case 2
CFA/IRT Model Predicted Observed Score
Same Predicted Subgroup Mean-levels
Different Predicted Subgroup Variances
Subgroup 1
Subgroup 2
Same Latent Score Distributions for each Subgroup
Latent Score
MLMB: Case 2
Machine Learning Model Predicted Score
Same Predicted Subgroup Mean-levels
Different Predicted Subgroup Variances
Subgroup 1
Subgroup 2
Same Ground-Truth Score Distributions for each Subgroup
Observed Score (Ground-truth score for ML model)

Figure 2. Measurement bias and machine learning measurement bias

Note: Measurement bias (MB) and Machine Learning Measurement Bias (MLMB) Case 1 represents a non-compensatory bias that creates different predicted subgroup distributions despite the same underlying subgroup distributions; MB and MLMB Case 2 represents a compensatory bias that creates equivalent predicted subgroup distributions although there is measurement bias.

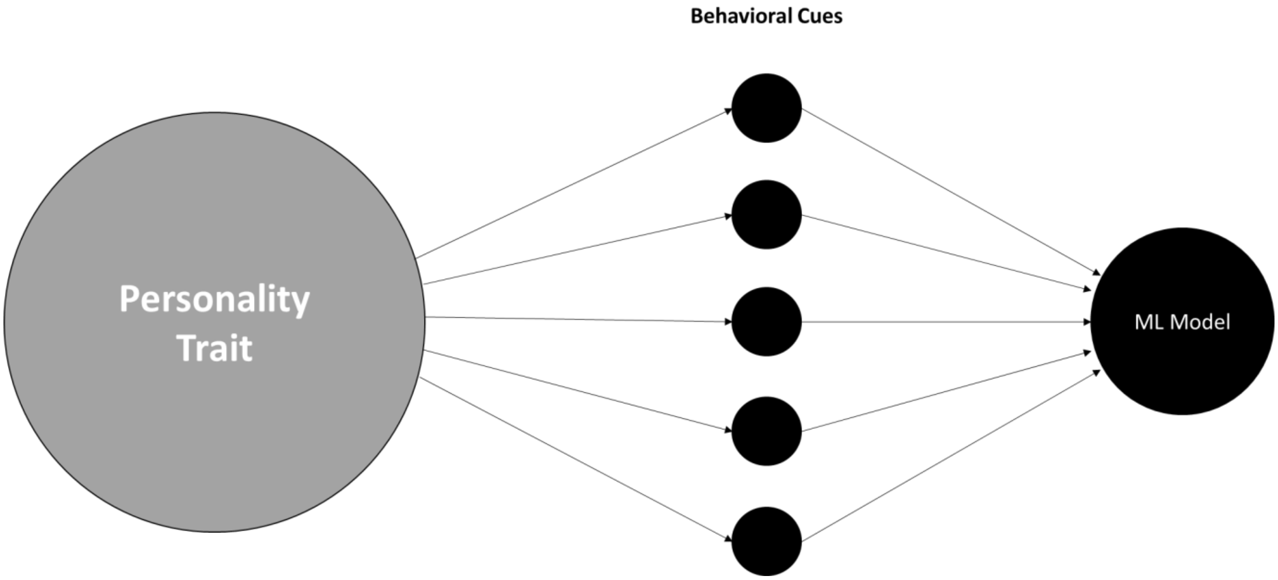
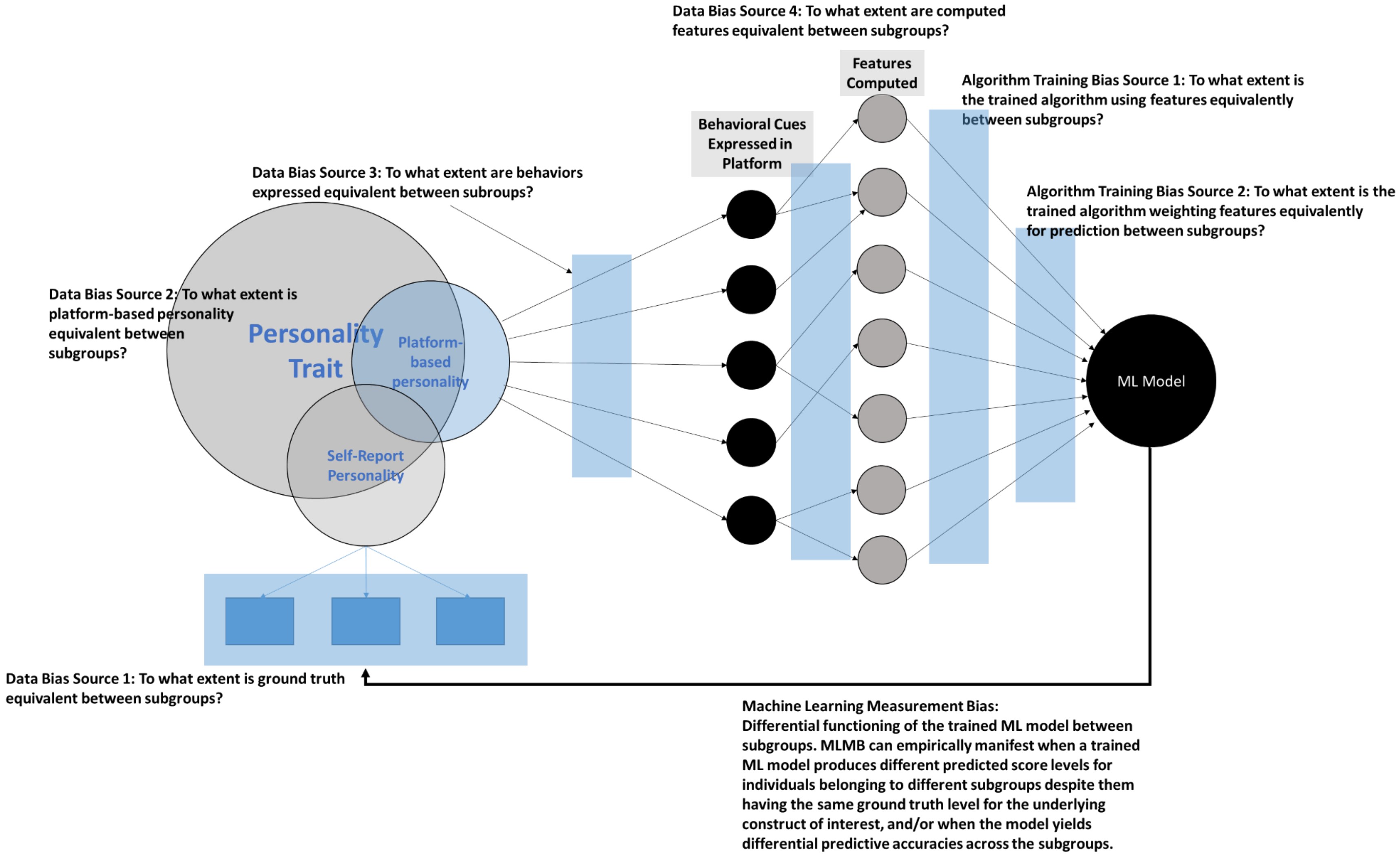

Behavioral Cues
Personality Trait
ML Model

Figure 3. The Brunswik lens model: Linking ML model to behavioral cues


Personality Trait
Platform-based personality
Self-Report Personality
Behavioral Cues Expressed in Platform
Features Computed
ML Model

Data Bias Source 4: To what extent are computed features equivalent between subgroups?

Algorithm Training Bias Source 1: To what extent is the trained algorithm using features equivalently between subgroups?

Algorithm Training Bias Source 2: To what extent is the trained algorithm weighting features equivalently for prediction between subgroups?

Data Bias Source 3: To what extent are behaviors expressed equivalent between subroups?

Data Bias Source 2: To what extent is platform-based personality equivalent between subgroups?

Data Bias Source 1: To what extent is ground truth equivalent between subgroups?

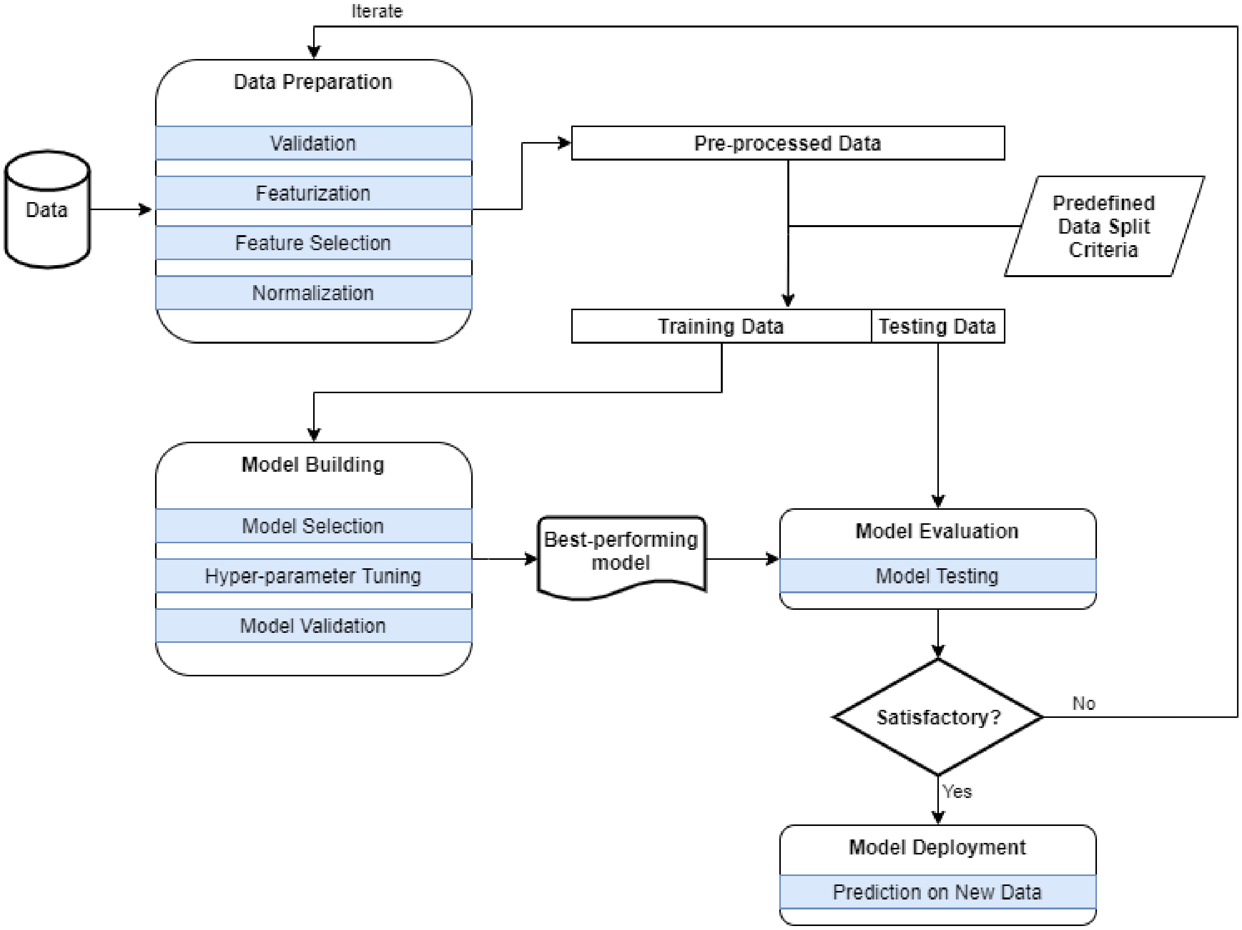
Machine Learning Measurement Bias: Differential functioning of the trained ML model between subgroups. MLMB can empirically manifest when a trained ML model produces different predicted score levels for individuals belonging to different subgroups despite them having the same ground truth level for the underlying construct of interest, and/or when the model yields differential predictive accuracies across the subgroups.

Figure 4. Expanding the Brunswik lens model to identify the sources of machine learning measurement bias: An illustration using personality as the focal construct

Note. Areas highlighted in blue represent possible sources of machine learning measurement bias; "platform-based personality": the personality construct measured by input data (e.g., online personality assessed by social media data) used in ML models to predict self-report personality

Appendix

This figure depicts a classic machine learning procedure designed to produce a trained model capable of making predictions on future data samples. Here we describe the stages and workflow with examples.


Iterate
Data Preparation
Validation
Featurization
Feature Selection
Normalization
Data
Pre-processed Data
Predefined Data Split Criteria
Training Data
Testing Data
Model Building
Model Selection
Hyper-parameter Tuning
Model Validation
Best-performing model
Model Evaluation
Model Testing
Satisfactory?
No
Yes
Model Deployment
Prediction on New Data

Data Preparation: In this first stage, the data samples are gathered from the database and prepared for modeling. In this stage, the data is checked to ensure that the data samples are valid. Missing or invalid data may be dropped from further processing or filled in with more reasonable values. Data that exists in a raw form (e.g., a video) may be converted to a set of features, a set of transformations of the raw data, which are more relevant to the modeling process (e.g., facial expression features derived from the video). Sometimes this set of features will contain redundancies (e.g., pairs of features that have a high correlation), so a feature selection process will eliminate redundant features so the later stages will have less data to deal with. Finally, the features will often be normalized (e.g., using z-scoring or min-max norming) so that the features are initially given a roughly equivalent potential to influence the outcomes once the model is built.

Data Splitting: The resulting set of featurized and normalized data are split into two sets of data: the data that will be used to build the model and the data that will be used to evaluate the model. The former set is typically called the *training data set,* and the latter the *test data set*. The split criteria are often determined without peeking at the data so as not to introduce any selection bias.

A common strategy is an 80/20 split, where 80% of the data is used for training and 20% is used for testing. Sometimes, randomized stratified sampling may be used to ensure that roughly equal proportions of similar data samples appear in both the training and test data sets, which improves the evaluation of the model's generalizability.

Model Training: In this stage, the training data set is used to build a model which is expected to make accurate predictions on the test data set. Researchers will decide which class of models they wish to use (e.g., linear least squares regression, polynomial regression, random forest, deep neural network). This decision may come from the successes of related literature, from experience, or researchers may elect to test multiple model classes on their problem. For each model class, a refinement procedure called hyper-parameter tuning is used to find values of hyper-parameter values (e.g., the regularization coefficients in Elastic Net models) that fit the data training data set samples the best. To help ensure the chosen hyper-parameters will help the model class generalize to unseen data, cross-validation may be employed. During cross-validation, the train data set is often randomly split into either 5 or 10 equal pieces or "folds" where, in turn, all but one fold is used to tune the hyper-parameters, then the chosen ones are used to evaluate the model on the remaining fold. Researchers may obtain summary statistics of accuracy (e.g., mean, variance) over the folds and compare these for different hyper-parameter settings to decide which hyper-parameters to choose. Finally, once the best hyper-parameter values are decided for a given model class, that model is *trained* by fitting the model to all the training data samples using the hyper-parameters to produce the best-performing model.

Model Evaluation: The trained model is evaluated using the test set data. The model produces a separate prediction for each test set data sample, and the prediction is compared to the known true outcome. The comparisons are aggregated over the test set samples providing a performance score (e.g., mean accuracy, F1 score). Since the model was trained on data samples independent of the test data set, the performance score is expected to generalize to future data samples.

Model Deployment: If the performance score is not high enough, researchers may repeat the entire process, changing prior decisions made during data preparation and model building to try to build a better model. On subsequent iterations, the same predefined split criteria are used to divide the data to avoid introducing any bias. If the model's performance is deemed acceptable, then it may be deployed and used to make predictions on new data samples.

Traditional ML vs. Deep Neural ML

The discussions in our paper focus on a traditional form of machine learning. In contrast, deep neural learning or deep learning (Goodfellow, Bengio, & Courville, 2016; Le Cun, Bengio, & Hinton, 2015) is a different class of models that have gained prominence over the past decade. These models are constructed by combining multiple 'layers' of artificial neural networks, consisting of an input layer, one or more 'hidden' (intermediate) layers, and an output layer; where hidden layers learn useful intermediate representations of the data by combining input features. For example, a deep learning algorithm will directly process raw images rather than features (e.g., textures, color) extracted from the images because it is sufficiently powerful to learn the features themselves. These deep neural networks have resulted in incredible performance improvements across various domains such as computer vision and natural language processing (Devlin, Chang, Lee, & Toutanova, 2018; Szegedy, Ioffe, Vanhoucke, & Alemi, 2017). They are also capable of representation learning in that they can learn features themselves by extracting patterns from raw data instead of requiring pre-specification of features like in the other modeling approaches. Deep learning models are incredibly complex with the number of free parameters in the tens to hundreds of thousands, so they require copious amounts of training data, and their interpretability is low. For these reasons and because most ML applications in psychology focus on traditional ML, we do not discuss deep learning specifically.

Glossary for Different Types of Machine Learning Models

Term	**Definition**	**Psychological Assessment Example**
Classification	Where the machine learning model's output is a discrete value. Binary classification tasks involve just two outcome options, but multi-class classification can involve many outcomes. Multiple binary classifiers can be used as an alternative to multi-class classification, and doing so allows each observation to be classified as including multiple emotions.	Min et al. (2021) classified whether tweets contained one or more of six discrete emotions.
Regression	Where the machine learning model's output is a continuous value, such as when predicting Big Five trait scores. There are different regression techniques (e.g., linear regression, ridge regression, lasso regression, etc.)	Park et al. (2015) inputted principal components from text data into ordinary least squares regression to generate continuous scores for the Big Five traits.
Decision Tree	A hierarchical "tree-like" model used to classify a data sample based on a sequence of (typically) univariate predictor tests. The outcome of each test either determines which test is performed next or results in a decision output (e.g., a class).	Sajjadiani et al. (2019) used decision trees (among many other algorithms) to classify whether applicants reported leaving their previous jobs due to involuntary turnover, avoiding bad jobs, or approaching better jobs. In their case, a single decision tree outperformed a random forest.
Random Forest	A model used to regress or classify a data sample based on the combined outcomes from multiple (i.e., a forest of) decision trees, where each decision tree is built using a random subset of the available predictors and observations, such that each consists of unique sequences of predictor tests.	Gladstone et al. (2019) used random forest to develop models that scored the Big Five traits from credit card spending records.
Support Vector Machines	A model used to regress or classify a data sample by developing a linear function that is used as a decision boundary. During model training, the linear function is carefully chosen to separate binary labeled data as accurately as possible while also maximizing the minimum distance to the nearest sample point on either side of the boundary. This model can be extended to perform multi-	Putka et al. (2018) used support vector machines (among many other algorithms) to compare the power of psychological scale scores vs. scale items in predicting job performance.

	class classification or regression or to certain types of nonlinear decision boundaries.	
Neural Networks	A model used to regress or classify a data sample through a series of (typically nonlinear) transformations of predictor values (known as layers). Transformed predictors can be combined and/or transformed again and passed on to future transformations (layers) in the series. The last transformations in the series often result in a small number of values that are either combined (for regression) or passed through a simple decision function (for classification).	Başaran & Ejimogu (2021) used Facebook activity, status, and demographics to predict Big Five personality traits via a neural network model.
Deep Neural Learning Models (Deep Learning)	A type of neural network where the number of (typically nonlinear) transformations in the series is large (at present, any quantity larger than one transformation is often considered “deep”).	Transformer-based transfer learning language models are deep neural networks. Min et al. (2021) fine-tuned BERT (Devlin et al., 2008) for emotion classification to estimate the public’s response to work from home orders during the COVID-19 pandemic.

Appendix References

Başaran, S., & Ejimogu, O. H. (2021). A Neural Network Approach for Predicting Personality From Facebook Data. *SAGE Open*, *11*(3), 21582440211032156.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological science*, *30*(7), 1087-1096.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Min, H., Peng, Y., Shoss, M., & Yang, B. (2021). Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. *Journal of Applied Psychology, 106*(2), 214–229.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., . . . Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934-952.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689-732.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207-1225.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on Artificial Intelligence*.