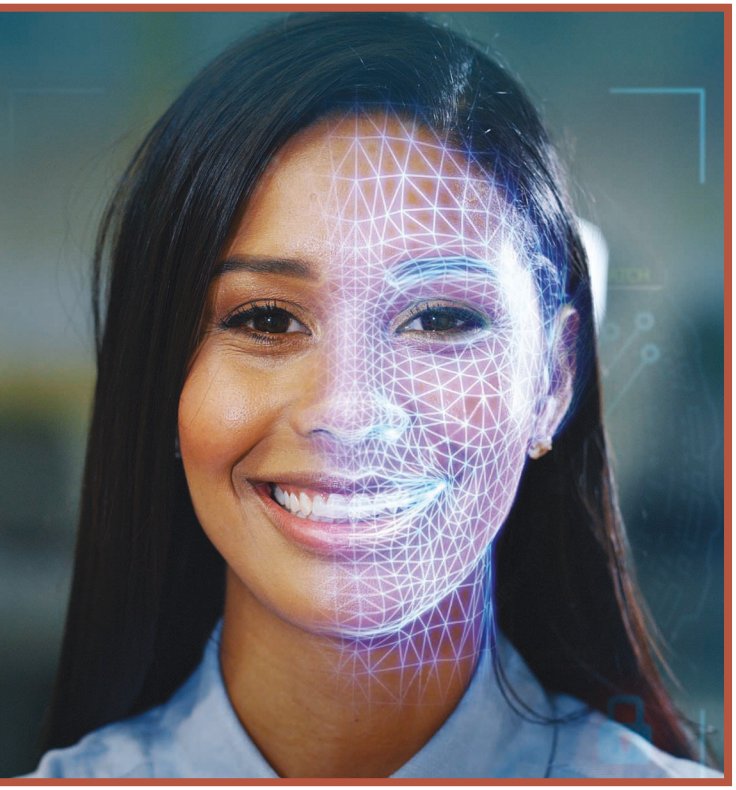
Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj,
Louis Tay, Sang Eun Woo, and Sidney K. D'Mello

Integrating Psychometrics and Computing Perspectives on Bias and Fairness in Affective Computing

A case study of automated video interviews



©SHUTTERSTOCK.COM/HQUALITY

Digital Object Identifier 10.1109/MSP.2021.3106615
Date of current version: 27 October 2021

We provide a psychometric-grounded exposition of bias and fairness as applied to a typical machine learning (ML) pipeline for affective computing (AC). We expand on an interpersonal communication framework to elucidate how to identify sources of bias that may arise in the process of inferring human emotions and other psychological constructs from observed behavior. The various methods and metrics for measuring fairness and bias are discussed, along with pertinent implications within the U.S. legal context. We illustrate how to measure some types of bias and fairness in a case study involving automatic personality and hireability inference from multimodal data collected in video interviews for mock job applications. We encourage AC researchers and practitioners to encapsulate bias and fairness in their research processes and products and to consider their role, agency, and responsibility in promoting equitable and just systems.

Introduction

The tools used in AC, which enable machines to identify people's behaviors and mental states, are being increasingly utilized in education, health care, and the workplace. One application is to aid in the allocation of limited resources (e.g., counseling, mental health care, in-person interviews) via automated screening [1]–[3]. In these types of high-stakes scenarios, the assessments provided by AC systems can directly affect the decision-making processes, which influence the amount of attention, care, and opportunities afforded to individuals. As such, it is important that these processes are accurate, unbiased, and fair because any deficiencies or errors present in these systems stemming from the data they were trained on, the types of algorithms used, or the decision-making processes themselves may disproportionately impact different groups of people and lead to ethical and legal concerns, not to mention pain and suffering for the vulnerable groups impacted. Simply put, AC systems must deter, not propagate, extant systems of inequity and injustice.

Fortunately, we have decades of guidance on how to construct fair and unbiased measurement systems.

 1053-5888/21©2021IEEE

The fields of educational and psychological measurement (i.e., psychometrics) have well-established, distinct definitions for test bias and fairness [4]. Great research progress is being made toward ethical data representations for artificial intelligence systems [5] and fair emotional expression recognition systems [6], yet most AC research ignores psychometric aspects entirely and, when considered, many studies of algorithmic bias treat the notions of bias and fairness somewhat interchangeably (e.g., [7]). Thus, a crucial first step toward reducing the potential short- and long-term disparities of AC systems is forming a consistent understanding of these terms. Accordingly, this article aims to provide an exposition of bias and fairness from a psychometric perspective, to ground these terms in a typical AC ML pipeline, and to enable AC researchers and practitioners to understand how sources of bias and unfairness contribute to observed manifestations or measurements of bias and unfairness.

The fields of educational and psychological measurement (i.e., psychometrics) have well-established, distinct definitions for test bias and fairness.

Our contributions are as follows. First, we define the psychometric meaning of bias and distinguish it from fairness, providing examples of each. Second, we present a typical ML pipeline used in AC to generate predictions for mental constructs (e.g., emotions) from physiological and behavioral data and decompose it into a recurrent sequence of information exchanges. We demonstrate that by representing these exchanges as noisy communication models; borrowed from classic information theory [8], one can identify possible sources of bias and unfairness at multiple stages in the pipeline. Third, we connect measurements of bias and fairness from recent computer science (CS) research to the psychometric definitions of bias and fairness. Finally, using automated pre-employment screening, or personnel selection, as an application domain, which utilizes many analytical tools from AC, we empirically demonstrate the process of testing for some types of bias and unfairness in automatic personality and hireability inference from video interviews.

Bias, fairness, and ML in AC

The terms *bias* and *fairness* are sometimes used interchangeably in reference to discrimination, and it is important to distinguish the two. Indeed, discrimination serves as an umbrella (legal) term encompassing both bias and fairness concerns [9], but these terms have distinct meanings that should not be confused.

The Standards for Educational and Psychological Testing (hereafter, the "Standards") has provided guidance on the development of valid, fair, and unbiased measurements since the first edition was released in 1966. In general, the Standards provides counsel for assessments (including computational ones) of psychological constructs intended to differentiate individuals, such as for mental health treatments or educational and employment opportunities. In AC, we are often interested in measuring latent constructs (i.e., an individual's states or traits) generally not directly observable, such as emotion, depression, and personality. In psychometrics (i.e., the study of psychological measurement), these constructs are measured using carefully crafted and validated assessments, including test items with correct/incorrect responses (e.g., intelligence tests), questionnaires with Likert-type scales, and other measurements (e.g., observations). In AC, these assessment items are replaced with automated inference from behaviors often obtained using cameras, microphones, and various physiological sensors. A typical ML pipeline for predicting a latent mental state involves passing data (e.g., behavioral observations about a person) through a trained ML model to obtain a prediction, which can later be used to make the decisions that affect people (e.g., to hire or not hire). Figure 1 illustrates this sequence of events and also depicts the different types of bias and fairness and their regions of concern with respect to this pipeline.

Fairness is a subjective perspective on the appropriateness of the way a construct is measured, how the measurement is used for decision making, and the explanations related to the use of the construct. Bias is any systematic error that differentially

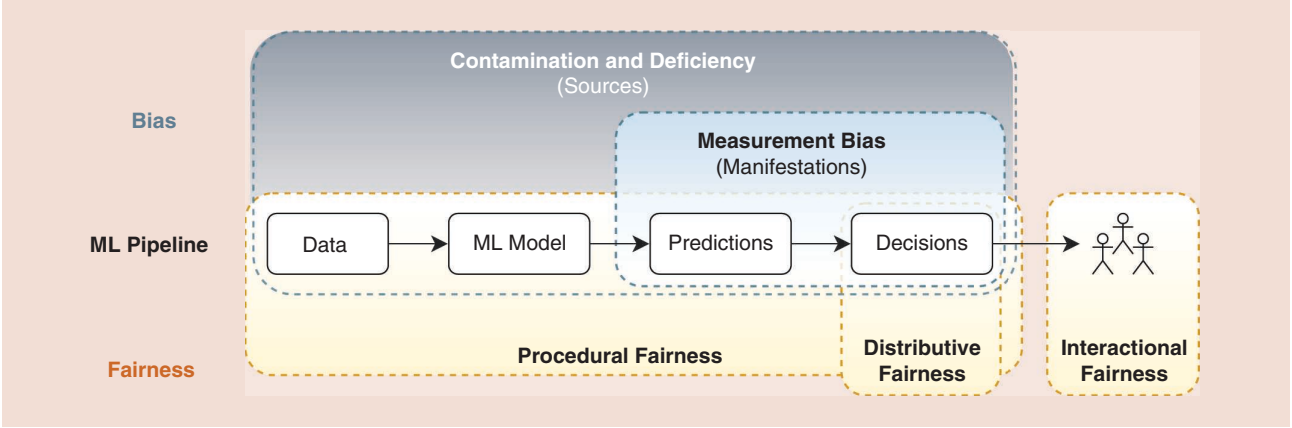


FIGURE 1. Different types of bias and fairness and their regions of concern with respect to a typical ML pipeline used for decision making, where the outcomes affect real people.

affects assessments of distinct groups of people. These are two different notions, but we often hear about them together because they both pertain to potential discrimination and the quality of decisions. The Standards considers bias to be subsumed by fairness in that a biased measurement is likely to be unfair. Yet, not all measurements viewed as unfair are biased, not all unbiased measurements are considered fair, nor will everyone view a biased measurement (and the subsequent decisions made using it) as unfair. These terms are sometimes conflated in CS and ML literature (e.g., [7] and [10]), but psychometrics offers a clear and established perspective on these topics.

Fairness

Fairness has no universal definition as it is a social, not psychometric, concept rooted in value judgments [11]. Fairness is a subjective evaluation (e.g., justice and morality), varying across cultures and societies, and in the context of organizations such as schools, hospitals, or corporations, organizational justice has been the predominant theoretical concept used to recognize perceptions of unfairness [12]. Although AC is not broadly tied to understanding people within organizations, examining fairness through this lens is highly illustrative of some of the difficulties and inherent tradeoffs (compare [13]) in fairness considerations within AC.

Organizational justice involves three key dimensions: distributive, interactional, and procedural fairness [12]. Distributive fairness regards the perceived fairness of outcomes and allocations of important resources (e.g., jobs). Interactional fairness regards how people perceive the explanations, rationales, and justifications for organizational decisions and how they perceive the interpersonal treatment they receive along the way. Procedural fairness regards the perceived fairness of the elements of the decision-making process. Procedural fairness is emphasized in the Standards because it is crucial that an assessment (e.g., ML predictions) does not generate different scores among subgroups if they have equivalent true scores. However, if there are differences among groups due to societal structures or biology, the assessment should accurately assess any potential differences. For example, a measurement of height should not show equal heights for men and women just to be "fair."

Procedural fairness is emphasized in the Standards because it is crucial that an assessment does not generate different scores among subgroups if they have equivalent true scores.

Each of these types of fairness is relevant in the context of AC research, tools, and products. For example, facial recognition and expression software has been a core component of the AC tool kit and used to gain insights into the expressed emotional dynamics during social interactions. This capability is being incorporated into ML systems that, for example, observe the expression dynamics of individuals in recorded video interviews to make inferences about personality and other potentially relevant characteristics for employment [1]. However, one well-known issue is that the underlying facial recognition software tends to be less accurate for Black individuals compared to White individuals [14]. In this context, distributive fairness is concerned with ways to enhance the equality of scores and outcomes for measurements, which include facial recognition. Interactional fairness would be concerned with enhancing the explainability of the ML pipeline decisions and seeking to provide acceptable justifications for them. Procedural fairness would be concerned with the use of (or error associated with) facial features for expression recognition, which may be indicative of group membership (e.g., skin color [14] or facial structure).

In the United States, laws and case law (Title VII of the Civil Rights Act of 1964; Age Discrimination in Employment Act of 1967; Americans with Disabilities Act of 1990; Civil Rights Act of 1991; Bostock v. Clayton County Georgia) clearly define the groups that are protected from employment discrimination: age, disability, race, religion or belief, sex, gender, lesbian, gay, bisexual, transgender, queer, and pregnancy or maternity. The U.S. Civil Rights Act of 1991 established that direct or indirect measurements of these group attributes cannot be used in the decision-making process for employment. This precedent establishes a hard line for procedural fairness for any automated system deployed within the United States and used to aid in employment decisions (other countries may have different restrictions). By extension, this means that facial expression recognition software used to aid in employment decisions in the United States cannot attempt to correct for its poorer performance for darker skin tones by being aware of skin color. Thus, these systems must remain group unaware (i.e., "fairness through unawareness") while also meeting the growing demands for fair outcomes (i.e., distributive fairness) and explainability (i.e., interactional fairness). Attaining fairness in these types of systems is a difficult and complex task, both from an engineering and social perspective.

This challenge becomes even more apparent when examining recent work on ML fairness. Many of these works emphasize distributive fairness, which is highly desirable, but it is often difficult to achieve because of inherent tradeoffs among differing perspectives on what is considered fair [13]. One perspective on distributive fairness is that of equality, or the notion that each person or group of persons receives the same outputs (e.g., job offers). Another perspective is equity, where each person or group of persons should receive outputs proportional to their inputs (e.g., more job offers go to those who demonstrate merit). A third perspective is that of need, or the notion that each person or group of persons receives outputs according to their necessity (e.g., persons lacking money or who are otherwise disadvantaged receive more job offers). These methods for distributing opportunities are fundamentally opposed, but each may be deemed fair according to individual differences in outlook. Creating fair ML systems amounts not only to transparency and measuring fairness but also a social buy-in from the organizations utilizing them and the stakeholders affected by them.

Bias

The term *bias* is semantically overloaded and has many specialized definitions in different contexts. Many are familiar

with bias in the form of implicit and explicit bias, which involve systematic errors of judgment among humans due to the demographic characteristics of a given target (e.g., race, gender, or religion) [15]. These types of bias relate to the systematic influences that alter human behaviors or judgments about others as a function of their group membership.

In psychological assessment, which characterizes much of the AC applications in this area, *bias* refers to any systematic error in a test score that differentially affects the performance of different groups of test takers [16, p. 23], where group membership is determined by distinguishing characteristics among the agents (e.g., gender or age). For example, any facial recognition software whose accuracy scores vary by race or gender would be biased. This is the definition of bias we adopt for the remainder of this article, and first distinguish between sources of bias and evidence or manifestations of bias.

The sources of bias in an assessment of a construct can broadly be attributed to either construct contamination or deficiency [4], as illustrated in Figure 2. *Contamination* refers to the sources that introduce construct-irrelevant variance, while *deficiency* refers to the omission of construct-relevant variance. If these types of errors universally inflate or deflate scores independent of group status, then the assessments may be described as inaccurate, but they would not necessarily be biased. Psychometric bias regards errors that differentially affect members of one group compared to members of another. For example, an AC system for judging hireability from tone of voice and nonverbal behaviors (ostensibly signaling competence) trained exclusively on White individuals (i.e., population bias, sampling bias, and representation bias) will tend to be deficient when assessing hireability patterns for other racial groups, and it will also be contaminated with the behavioral patterns applicable to whites but not other racial groups.

Contamination refers to the sources that introduce construct-irrelevant variance, while _deficiency_ refers to the omission of construct-relevant variance.

The manifestation of bias in ML, or the bias that we observe, mostly comes from measurement bias. Measurement bias occurs when assessment scores contain systematic error that is not relevant to the construct of interest, such as an ML pipeline producing predictions of personality scores, which are systematically influenced by race. In this case, measurement bias would be observed if racial subgroups have the same ground-truth scores but the assessment systematically provides different scores.

Importantly, we do not make any intentional attributions to bias, instead arguing that it arises from the involvement of humans in the process. For example, in AC, we typically rely on human-produced (i.e., self- or other-reported) assessments of constructs to serve as ground-truth labels for ML modeling. This step is necessary because the constructs of interest are latent, meaning they are hidden and cannot be directly observed. Intuitively, we may imagine that differences in the mean-ground-truth label for different groups indicates bias, but the Standards states that group differences in outcomes do not in themselves indicate that a testing application is biased or unfair [4, p. 54]. Any differences in group means may reflect true differences in the underlying construct. Simply put, a measurement of height that systematically indicates that men are on average taller women is not biased.

Ground truth in AC is sometimes obtained with the aid of validated self-reported psychological measurements, which have ostensibly been tested for bias with various subgroups. However, when observers are used to obtain ratings or annotations, steps should be taken to ensure ground-truth validity, including conducting frame-of-reference training [17], using a panel of diverse annotators, monitoring annotation quality (e.g., via interrater reliability/agreement), and removing outlying or low-quality annotators. These kinds of steps result in a collection of annotations, which can be better trusted in aggregate as accurate ground-truth representations, where any group differences are a reflection of true differences among groups rather than bias. These steps are absolutely essential for bias analysis—if human implicit/explicit bias contaminates the ground-truth measurement, then the resulting ML assessment is very likely to reflect these biases.

Identifying Sources of Bias

We endeavor to provide a framework for identifying the possible sources of bias in AC ML (Table 1) by deconstructing the ML pipeline into a recurrent sequence of exchanges of information among different pieces of the pipeline. We can then examine the sources of bias associated with each piece to understand how bias emerges, propagates, and manifests at various points throughout the ML process. To do so, we expand upon a common conceptual framework employed in AC and natural language processing to understand noisy information exchange [8], [18]: the two-agent communication model. We caution, however, that any list of biases, such as the examples presented in Table 1, only serve as a reference for researchers and may not be comprehensive

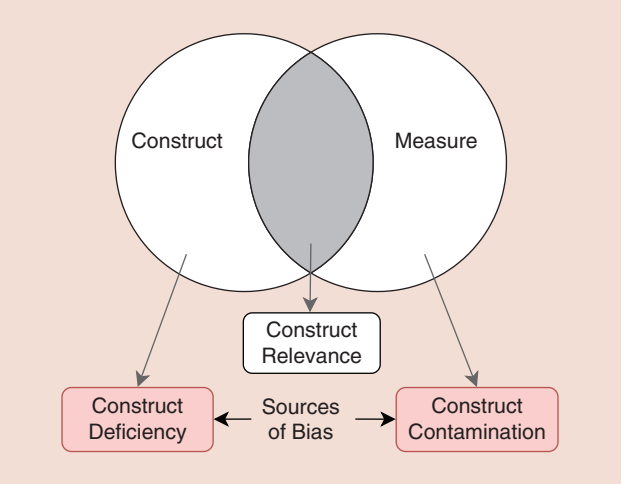


FIGURE 2. The sources of measurement bias in construct assessment [4] (reproduced from [3]).

enough to represent all the possible sources of bias in any given study.

Communication model for bias identification

Mehu and Scherer [19] helped bridge social signal processing with psychology and ethology by considering how the (un)reliable and contextual nature of human behavior impacts communication and efforts to automate its understanding. In this work, they considered communication as an encoded exchange of information, hearkening back to early communication models such as the Shannon–Weaver one [8], first proposed in 1948, or the sender-message-channel-receiver model, later introduced by David Berlo [18]. These models consider communication as a process (not necessarily serial) among a sender, who encodes a message; a receiver, who perceives and decodes the message; and a channel, through which the message passes between the two, as illustrated in Figure 3.

Table 1. The sample sources of bias relevant to ML for AC.

	Bias Term	Meaning
Deficiency	Selection/sampling	Statistics, demographics, and user characteristics are different in the user population than in the collected data
–	Omitted variable	One or more important variables are left out of the model
Contamination	Historical	Existing systemic biases seep into the data collection process
–	Representation	Decision makers incorrectly apply priors from an earlier situation they perceive as similar to the current one
–	Behavioral	Behavior in the user population differs from behavior in the training data
–	Presentation	When the order or style of information presented to participants causes faulty reasoning or alters their behavior
–	Observer	The tendency for people to subconsciously project their expectations onto their observations

Note: Although these terms are commonly dubbed *bias* and may cause psychometric bias, they are not themselves *psychometric bias* (i.e., a systematic error differentially affecting groups).

The information in this diagram flows from left to right, starting with a source concept, representing what the sender intends to communicate. This concept is then encoded as a sequence of behavioral actions (the action plan). These actions may include speaking, gesturing, touching, typing, or generally any form of sensory output. The speaker attempts to execute the actions to try to express the source concept to a receiver by means of a communication channel, such as air (carrying vocalizations) or a digital video recording. This channel is considered to be a noisy information tunnel where the encoded message may be altered on its way to the receiver and may thus interfere with the receiver's interpretation and understanding. The receiver perceives a potentially contaminated or deficient version of these expressions (e.g., via sensory inputs) and generates an internal representation of the expression. Finally, this representation is decoded to form the receiver's estimated concept, a version of the sender's source concept. Individual differences (e.g., experiences, behaviors, or genetic traits) in the sender and receiver may separately influence their encoding and decoding of the message which, together with noise in the communication channel, are potential sources of bias.

To make this model more concrete, consider an example relevant to AC where the sender is a speaker and attempts to describe her feelings (source concept) about a recent event to a friend. The speaker recalls her prior experiences to decide (action plan) how to describe her feelings. Her voice travels through the air (communication channel) to the receiver, who perceives the utterances and forms an internal representation of the words. The word representations are decoded to form the receiver's estimated concept and give meaning to the message.

In this example, the speaker's and receiver's individual differences as well as the communication channel may be sources of bias. If the speaker has a speech impediment (speaker individual difference) that momentarily interferes with her vocalizations, then there may be bias in the form of a deficiency in the information exchanged. Suppose the receiver is elderly and suffers from high-frequency hearing loss, then words may be lost during perception (communication channel). An elderly receiver may also have more trouble remembering the entirety of the message (receiver individual differences). If we assess the

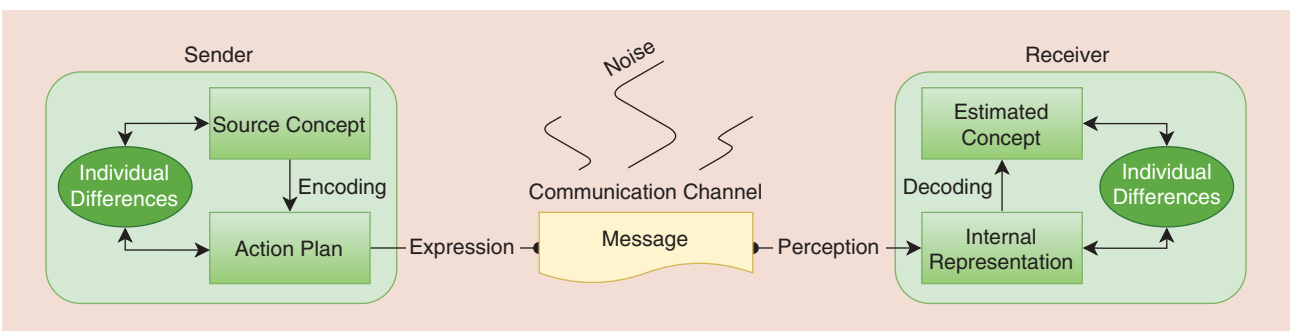


FIGURE 3. A one-way communication model for the transmission of a source concept (information) sent by one agent (sender) and received by another (receiver). Successful communication in this framework relies on the proper encoding and decoding of a concept—moderated by each agent's individual differences—and also on minimal interference (noise) from the communication channel. The agents' individual differences and communication-channel noise represent potential sources of bias.

accuracy of this information exchange based on a successful transmission of the sender's source concept to the receiver, then this type of information exchange would be biased as it is systematically less accurate for people with speech impediments, hearing loss, or memory difficulties.

The model presented in Figure 3 enables us to enumerate the potential sources of bias during information exchange: the sender's individual differences, systematic noise in the communication channel, and the receiver's individual differences. Referring back to our notions of contamination and deficiency from the Standards, these bias sources can affect the information by contaminating it and/or facilitating omission. We can chain elements of this model together to understand how biases influence communication in larger systems.

ML as a communication process

Let us examine a typical process for training a supervised ML model for AC. First, data (face, voice, physiology, and actions) are collected from a group of participants, usually selected out of convenience, while they engage with a stimulus (stimuli) and complete a task (tasks). Then, for each participant, a set of labels of subjective constructs (e.g., emotions) are collected using self-annotation or a panel of human observers. These assessments are combined or fused (e.g., by averaging) to form a ground-truth representation. Separately, a machine observer generates a set of features representing the participants' external behaviors and internal (e.g., physiological) responses. A model is trained using the machine-observed features and the ground-truth scores and subsequently tested using a predetermined evaluation metric, as defined and operationalized by stakeholders, to assess whether the model meets the goals of the project. Each of these steps may be repeated iteratively until the model is satisfactory.

Figure 4 illustrates this process for training AC ML models, beginning with a participant (left) and ending with satisfied stakeholders (right). This process includes several steps where information is exchanged between various agents, so we can utilize the noisy communication model from the previous section to represent the flow of information. Each information producer or consumer in the process is represented as a sender or receiver (green boxes). The information passed between them is represented as a message channel (yellow, wavy boxes).

Given this version of the ML process represented as a sequence of communication model exchanges, we can start to think about each step in the pipeline as being a potential source of bias. Just as before, each message channel may be subjected to external noise, which may differentially corrupt or omit information transmitted from the source sender. Likewise, the senders and receivers themselves may also introduce bias to the ML process when they send or receive messages based on their individual differences.

Taking the time to depict the flow of information in the ML process, whether it follows this typical AC workflow or not, enables stakeholders to produce a comprehensive set of bias sources from which construct-relevant information may be ignored (deficiency) or from which construct-irrelevant information may be added (contamination). However, the potential impact of contamination and deficiency from each source is not equally impactful or relevant in all settings. For example, the data produced by a machine observer are often communicated to the learning model digitally and may by temporarily stored as a file in computer memory. The noise influencing this digital communication channel is characterized by bit corruption and/or read/write errors, which are unlikely occurrences in modern robust computer hardware and even more unlikely to differentially impact groups. On the other hand, when interactions occur over videoconferencing, known deficiencies in the communication medium, such as bandwidth, which vary based on socioeconomic status, can be a source of bias.

The potential bias introduced by the ML model itself, taken as another example, illustrates the key distinction between the sources and manifestations of bias. Some ML models, once fully trained, result in a fixed and deterministic algorithm, which always maps the same feature inputs to a particular prediction (e.g., random forests, linear regressions, or neural networks). Once these models are deployed, they never change and thus cannot draw from prior predictions to introduce an additional bias (i.e., a bias other than the training data) into their predictions. These types of models therefore cannot be sources of bias themselves because they lack any agency and are not influenced by prior experiences (i.e., individual differences from Figure 3). The learning models that do modify future predictions based on inaccuracies in the past (e.g., online ML and reinforcement learning) have a memory and may appropriately be considered sources of bias. When differential outcomes across groups are observed in the output from the ML model (i.e., manifested bias) and the model itself cannot be a source of bias, then the bias source is upstream. It may be directly upstream (e.g., the machine observer), or it may stem from previous decisions made by the stakeholders (see Figure 4) who, for example, may have selected a machine observer that is not equally accurate across groups, as in the case of facial expression recognition software that is less accurate for people with darker skin tones, or eye trackers that have difficulty for those with corrective vision.

Measuring bias and fairness

Let us consider how to measure bias and fairness in practice. Many formalized definitions of bias and fairness have been proposed and explored in the ML literature. We refer interested readers to [10], [20], and [21] for an overview. Here we focus on metrics for measuring psychometric bias and fairness when ML models are used to measure psychological constructs such as emotion or personality.

Table 2 lists several bias and fairness metrics relevant to AC and categorizes them according to their inputs by pipeline stage, as illustrated in Figure 4. This list is not exhaustive and is only intended to be instructive to readers when considering how bias and fairness can be measured at different ML pipeline stages. Note that each of the three stages near the bottom of the figure (i.e., the feature, prediction, and decision stages) are represented in the table, but not in the ground-truth stage. This is because in AC, the constructs of interest are always

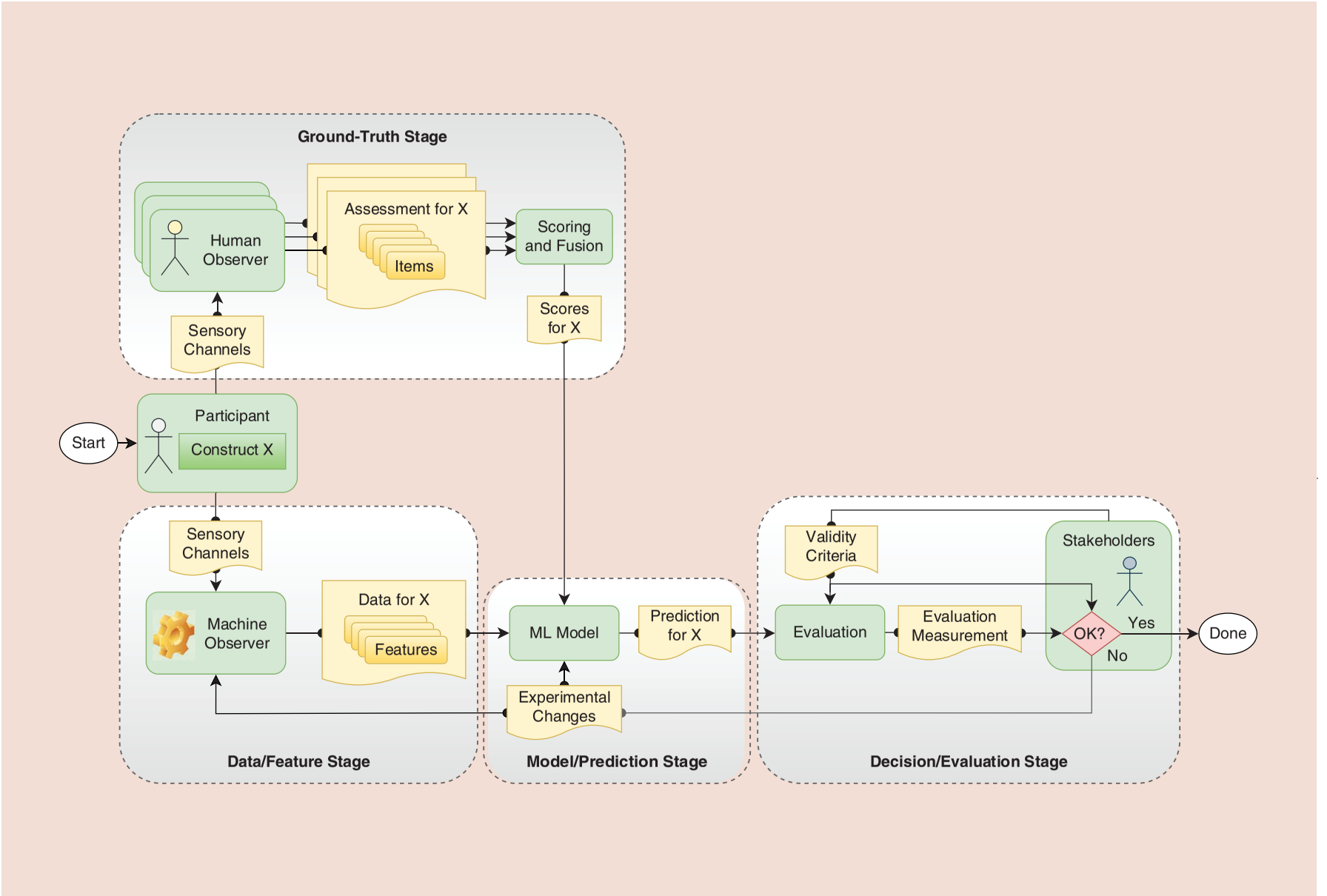


FIGURE 4. A deconstructed AC ML pipeline for developing a trained ML model is capable of predicting a participant's construct "X" (e.g., emotion). Noisy information exchanges (i.e., the yellow, wavy boxes) between each agent (i.e., the green box senders and/or receivers) in the pipeline are modeled using the noisy communication model (Figure 3). The elements are grouped into pipeline stages (similar to Figure 1) and labeled (gray dashed boxes).

latent (e.g., emotions), and thus a ground-truth measurement must first be established to evaluate bias. Any sources of construct contamination or deficiency in the ground-truth stage can be evaluated using traditional psychometric techniques (e.g., differential item functioning and differential prediction) and is not our focus here.

Bias metrics

Bias metrics come in a variety of forms and are designed to help probe for group differences (see the top half of Table 2). Each of these measurements is mathematically defined, but we omit the mathematical definitions for simplicity. Interested readers are referred to [10], [20], and [21] for more information, with our note of caution that the terms *bias* and *fairness* are sometimes used interchangeably in the CS literature.

Bias metrics at the feature stage are concerned with the contamination or deficiency of construct-relevant information contained within the predictors themselves. "Fairness through unawareness" is a binary metric that considers whether group membership is included as a predictor, which is often (but not always) construct irrelevant and should presumably be excluded to minimize bias. The other measurements of bias at the prediction and decision stages are tied to the accuracy of the models' predictions and are designed to check for unexpected differences in accuracy between groups. Intuitively, a prediction that is less accurate for one group compared to another contains systematic error, which disproportionately affects one group over another, whether that error is introduced via contamination or deficiency. These measurements can only provide evidence of manifested bias, so the bias source is always somewhere upstream of the measurement (see Figure 4).

The bias measurements that capture group differences in accuracy depend on the measurements of accuracy themselves. For example, studies that use correlational or mean-level measurements of accuracy can use group differences in these measurements as relevant measurements of bias. Similarly, for studies involving binary label prediction, accuracy measurements derived from the confusion matrix, such as treatment equality or equalized odds, can be relevant bias measurements. Bias can be further evaluated at the decision stage for some decision function by examining group differences in prediction-based outcomes when compared to the outcomes resulting from applying the same decision function to the ground-truth construct labels.

Once the bias metrics are computed, a separate question is how to interpret them. When the same bias measurements are reported in related research, direct comparisons can be made. However, in the absence of comparable bias measurements, differences in accuracy between groups can be difficult to interpret. One solution is to implement multiple ML prediction experiments with small changes, perhaps involving bias-reduction strategies, and then compare bias measurements (with accompanying statistical tests) to assess the effects of these changes.

Fairness metrics

Fairness is not concerned with the differential group accuracy but rather with how information is consumed and transformed to produce an actionable result as well as how different people are treated and impacted by the decisions. The bottom half of Table 2 presents some fairness measurements that are relevant in AC, but readers are referred to [10] and [20] for more comprehensive lists.

Table 2. Sample bias and fairness metrics relevant to ML for AC.

	Stage(s)	Name	Description
Bias	Features	Fairness through unawareness	Group membership is not used in an assessment
Procedural Fairness/ Measurement Bias	Prediction	Correlational accuracy	Equal correlations between prediction and ground truth across groups
—	—	Differential item functioning	Equal-item total correlations for annotations and/or predictions
—	Prediction/decision	Effect-size difference	Effect sizes between groups in predictions are equal to effect sizes between groups in ground truth
—	—	Treatment equality	Equal ratio of false negatives to false positives across groups
—	—	Equalized odds	Equal number of group true and false positive rates
—	—	Equal opportunity	Equal number of group true positive rates
—	—	Predictive equality	Equal number of group false positive rates
—	—	Overall accuracy equality	Equal number of confusion matrices across groups
—	—	Predictive parity	Chance of individual selection across groups is equal using ground truth and predictions
Distributive Fairness	Prediction/decision	Statistical parity/group fairness/ adverse impact	Equal number of group passing or hiring rates
—	—	AUC parity	Area under the receiver operating characteristic curve is equal across groups
—	—	Fairness through awareness	Equal number of predictions is given to similar individuals, given group knowledge
—	—	Counter-factual fairness	Equal number of predictions is given to individuals if hypothetically assigned to different groups
—	Decision	Conditional demographic parity	Decisions are independent of groups given the data
—	—	Single threshold	A single decision threshold is used for everyone

AUC: area under the curve. The metrics are categorized according to their inputs by pipeline stage as illustrated in Figure 4.

Each fairness measurement is based on a different interpretation of fairness and is therefore not necessarily compatible with other measurements of fairness. For example, counterfactual fairness assumes that predictions should be equal for an individual regardless of group membership, which presumes knowledge of group types and intentionally ignores true group membership for individuals. Conditional demographic parity suggests that decisions are fair when they are independent of the relevant groups, given a particular set of data, which assumes knowledge of group types and may or may not include true group membership for individuals. Fairness through awareness assumes that group membership contains useful information for adjusting predictions to make them more accurate and thus should be included in an assessment. Researchers and practitioners should be aware of the underlying assumptions imposed by each metric and whether they are compatible with stakeholder goals and legal restrictions.

Additionally, fairness metrics are distinct from measurement bias metrics and need to be considered separately for stakeholders to evaluate the benefits and potential harms of a deployed automated AC assessment tool. In a hypothetical experiment using real-world data, Kleinberg et al. [22] showed that an algorithm for deciding college admissions that is given knowledge about demographics (e.g., fairness through awareness) could help inform admissions committees in admitting a greater proportion of black and African American students (who are underrepresented in U.S. colleges) while also meeting or exceeding the average student body GPA goals. Notably, any algorithm that utilizes demographics to measure some construct is inherently biased because the algorithm is using construct-irrelevant information (e.g., race) to measure the construct. This is one perhaps counterintuitive example where increasing the bias of an assessment can lead to more fair outcomes.

Finally, some of the proposed measurements of fairness make strong assumptions about the true distribution of construct scores in the population across groups, which may or may not be reflected in the ground truth. For example, statistical parity, group fairness, and adverse impact (AI) are all concerned with the equality of acceptance rates across groups, presuming that ground-truth group score distributions should be equal. The Standards explicitly rejects these definitions of fairness because real differences may exist between groups (e.g., women tend to be perceived as slightly more extroverted than men [23]); however, it points out that group differences should cause additional scrutiny for other potential sources of measurement unfairness and bias. In certain high-stakes scenarios where automated AC assessment tools are used, such as employee selection in the United States, noticeably different hiring rates (AI) may constitute prima facie evidence of discrimination [9]. In spite of the strong (and sometimes perhaps inaccurate) assumptions made by these fairness measurements, it is crucial that researchers and practitioners engaged in developing high-stakes AC systems evaluate and use them to avoid any ethical or legal concerns and to avoid harming vulnerable populations.

Case study: Automated video interviews

We demonstrate the process of mapping and measuring potential sources of bias and fairness in a case study of automated video interviews (AVIs). In AVIs, job candidates are given a series of questions and asked to record their answers as part of a one-way (or asynchronous) interview. AVIs use computer software to ingest the recordings and generate behavioral features, which are inputted into ML models to score interviewee knowledge, skills, abilities, or other characteristics (e.g., personality) to help companies screen the candidates (e.g., a yes-or-no decision about whether to proceed with in-person interviews or hiring [1]). Human annotations are often used during the ML model development process as a ground-truth reference. Human assessments of these traits are based on the dynamics of vocalization, body expression, linguistic cues, perceived emotions, and other social signals as collected by speech and natural language processing, computer vision, and various other AC tools.

Fortune 500 companies are increasingly interested in utilizing AVIs to help screen job candidates more efficiently and effectively, but there has recently been push back due to potential biases in these systems [24]. For instance, in an in-person mock job interview experiment, Muralidhar et al. [2] observed that the automated assessment often rated males substantially higher than females on professional, social, and communication skills (e.g., enthusiasm, competence, and motivation), postulating that the differences were due to gender stereotyping in social cue perception while collecting ground-truth scores. It is both legally and ethically imperative that developers of these high-stakes AVI systems carefully analyze bias and fairness to avoid social harm and aid in promoting just systems. We demonstrate this process using a data set of mock video interviews and an ML model trained to make assessments of job-relevant traits.

Data and models

Description

A total of 511 college students (62% female, 37% male, and 1% nonbinary) were recruited to participate in a mock video interview for a hypothetical job opening. The participants were presented with six interview questions in random order, one at a time; and for each question, they were given a few minutes to prepare a response before recording a short video (1–3 min) of themselves answering it. A full list of the questions can be found in [3]; one sample question is, "Tell me about a recent uncomfortable or difficult work situation. How did you approach this situation? What happened?"

Ground truth

At least three members of a larger panel of trained human annotators, acting as the interview committee, rated the videos for each participant. The seven constructs of interest were the Big Five personality traits (agreeableness, openness, extraversion, emotional stability, and conscientiousness), perceived intellect, and hireability. The ratings were provided separately by each rater and were based on all the responses from a given participant. After establishing adequate interrater

reliability [a one-way, random, average intraclass correlation coefficient (ICC)(1,k) = .67)], the ratings from the annotators were averaged to generate a ground-truth score for each participant and construct.

Features

A set of features was extracted from each video's visual and audio channels, capturing verbal [e.g., *n*-gram (word and phrases) frequencies, linguistic inquiry, and word count categories], paraverbal [e.g., loudness, Mel-frequency cepstral coefficients (MFCCs), jitter, and shimmer], and nonverbal (e.g., facial action units and total body motion) behaviors. Unigram, bigram, and trigram features were computed from the audio transcripts produced by the IBM Watson automatic speech recognition service. Bigrams and trigrams with a point-wise mutual information (PMI) measurement of less than four were dropped to reduce the overall feature count (per [25]). For each participant, a set of statistical functionals was independently applied to the features extracted from each of the six videos, including median, standard deviation, minimum, maximum and range, and then averaged across the recordings to produce one multidimensional observation per participant throughout the entire mock interview.

ML model

A random forest learning model was selected to make predictions of the constructs based on the audio and video features (although any model would suffice for the following bias/fairness analysis). The data set was partitioned separately for each construct into five equally sized folds, utilizing a stratified sampling approach such that each construct's ground-truth distributions were roughly equal across folds. As each data sample corresponded to a unique participant, the folds were participant independent. Nested five-fold cross validation was used to tune the hyperparameters of the random forest algorithm (number of decision trees {10, 250, 500}, maximum depth {10, 50}) and the verbal feature extractor (stop words {none, English}, minimum term frequency {.01, .02, .03}). In total, there were approximately 7,877 features after PMI filtering, depending on the fold, construct, and the words uttered by participants, which comprised 250 visual, 125 paraverbal, and roughly 7,502 verbal features.

Bias and fairness results

We evaluate the bias and fairness of the AVI ML pipeline with respect to gender at the feature, prediction, and decision stages, in line with the stages illustrated in Figure 4 and mentioned in Table 2. Our data set contained only four participants with nonbinary gender affiliations, so we exclude them in the following analysis, noting that more data would be necessary to understand how bias and fairness concerns impact the excluded gender groups.

Feature stage

We adopt a fairness through unawareness strategy to minimize bias, where the gender of each participant is not included in the features used to train the ML model. By this definition, personality, intelligence, and hireability do not depend on gender information, so including gender would contaminate the ML predictions with construct-irrelevant information and introduce bias. Although omitting gender seems to satisfy this bias goal, we note that gender information is often encoded in other features such as vocal pitch, shimmer, and MFCCs, which we do include. These features likely contain both construct-relevant information and gender bias, so a careful analysis of the impact of bias would be necessary.

An exploration of bias-mitigation strategies is outside the scope of this article, but various techniques such as the exclusion of gender-biased features (i.e., features that carry a lot of gender-relevant information) [3] or fair representation learning should be considered and tested. Although the per-gender normalization of features may seem justifiable, the U.S. Civil Rights Act of 1991 explicitly outlaws using demographic information to indirectly adjust scores.

Prediction stage

We evaluate the ML model using a Spearman correlation, which examines the rank-order consistency between the predicted and ground-truth scores, a relevant metric for when applicants will be ranked against each other [26]. The left section of Table 3 shows the correlations for all participants, separately for women and men, and also the correlation difference between women and men, which provides one measurement of bias. By themselves, these measurements are difficult to interpret but would serve as a baseline for comparison in attempts to mitigate bias. Larger group correlational differences, such as the –.12 for extraversion or .11 for conscientiousness, arouse our suspicions as evidence of manifested gender bias and warrant further investigation.

The second section of Table 3 shows an *effect-size difference* (see Table 2) bias measurement operationalized using Cohen's d effect size [Cohen's $d = (\bar{x}_m - \bar{x}_w / s)$, where $s^2 = \left[[(n_m - 1)s_m^2 + (n_w - 1)s_w^2] / (n_m + n_w - 2) \right]$ m = men, w = women). The mean-effect-size differences between men and women are computed separately in the ground-truth labels and in the predictions, and then the difference in Cohen's d between the two constitutes a measurement of bias. The general guidelines for interpreting d dictate that values between .2 and .5 are considered small-to-medium-effect sizes, and thus we can see evidence of potential bias in the predicted values for all the constructs except agreeableness (for which the predictions were so inaccurate that invalidity is a bigger concern than bias). A further investigation reveals that the predicted values from the ML model range between approximately 3 and 6, while the ground truth ranges from 1 to 7. The restricted range of the predictions may be contributing to the larger differences in d (due to lower standard deviation) and warrant further investigation.

Decision stage

To assess compliance with the U.S. Civil Rights Act of 1964 [9], we focus on AI as a decision-stage fairness measurement. AI is defined as the *quotient of group selection ratios*, in our case computed by $\min[(SR_W / SR_M), (SR_M / SR_W)]$, where SR_W is the number of women accepted by some binary decision process divided by the total women applicants (likewise for

men). We simulate a realistic decision function by selecting the top k candidates among all participants based on the construct predictions, and then we set k equal to 10% of all participants so that we have a sufficiently large sample size for computing group selection ratios.

Taking hireability as an example trait and using the ground-truth scores as a baseline, we find the selection ratio for women is .1 and for men is .1, resulting in an AI ratio=1, which suggests that selecting for hireability in the ground truth is equitable. Using hireability predictions, the selection ratio for women is .11 and for men is .08, yielding an AI ratio of .7. In the United States, this violates the "four-fifths rule" (29 CFR§1607.4) and would be considered prima facie evidence to support a legal discrimination claim. Although the underlying cause of this unfairness may be bias upstream from the selection procedure, any system deployed for employment selection (at least in the United States) needs to demonstrate compliance with the four-fifths rule regardless of whether any bias can be found. Further exploration of the (un) fairness of other decision thresholds (i.e., other settings for k) should be conducted to assess the sensitivity of the AI ratio for this AVI system.

Discussion

This article presented a framework for understanding psychometric bias and fairness according to the Standards in the context of a machine-based assessment of emotion and its related constructs. We aimed to demonstrate that deconstructing a complex ML pipeline into a recurrent sequence of information exchanges and then treating those exchanges as noisy communication channels facilitates an understanding of how bias emerges, propagates, and manifests at various points throughout the ML development process. We suggest that decomposing other complex systems involving information exchange in this same manner will enable more prescriptive bias and fairness assessments.

We were not able to cover many important issues in this article and want to conclude with some remarks about future considerations. It has been recognized that there is often a tradeoff between creating the most accurate model possible and reducing bias or enhancing fairness [10]. Model validity, bias, and fairness are all crucial considerations for automated AC systems, and therefore, ongoing and future work should consider holistic optimization approaches rather than more traditional optimization of accuracy alone. More work is needed to investigate and normalize the methods that effectively maximize these outcomes, but some Pareto-optimization techniques already show promise [27].

In summary, the researchers and practitioners in AC developing the algorithms, software, and tools employed to aid in decision making have an ethical and moral responsibility to assess systemic errors and gauge the disproportionate impact that they can impose on people. We hope this exposition has been instructive in understanding, identifying, and measuring bias and unfairness, taking the first step toward this goal.

Researcher and practitioners in AC have an ethical and moral responsibility to assess systemic errors and gauge the disproportionate impact that they can impose on people.

Acknowledgments

This research was supported by the National Science Foundation (IIS 1921087 and IIS 1921111) and the National Science Foundation (NSF) National AI Institute for Student-AI Teaming (DRL 2019805). The opinions expressed are those of the authors and do not represent the views of the NSF. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Purdue University Institutional Review Board protocol IRB-2019-56, *Understanding Biases in Video Interviews.*

Authors

Brandon M. Booth (brandon.m.booth@gmail.com) received his Ph.D. degree from the University of Southern California and is currently a postdoctoral research associate in the Emotive Computing Lab at the University of Colorado Boulder, Boulder, Colorado, 80309, USA. His research focuses on using multimodal machine learning techniques to model human perception, behavior, and experiences, and developing algorithms to reduce the impact of inadvertent human biases and errors. He has a diverse industry background researching, publishing, and developing video games, serious games, robots, computer vision and human–computer interactions systems, and geospatiotemporal visualizers.

Table 3. The example bias and fairness measurements in our case study.

Construct	Spearman ρ				Cohen's d			AI Ratio	
	All	Women	Men	Women–Men	True	Pred	True/Pred	True	Pred
Agreeableness	.03	.01	.08	–.07	–.13	–.22	.09	**.77**	**.48**
Openness	.34	.37	.27	.1	–.13	–.39	**.26**	.92	**.32**
Emotional Stability	.31	.28	.21	.07	.36	.66	**–.3**	**.57**	**.31**
Conscientiousness	.33	.34	.23	**.11**	–.34	–.61	**.27**	**.36**	**.14**
Extraversion	.47	.42	.54	**–.12**	–.09	–.49	**.4**	.84	**.36**
Perceived Intelligence	.4	.39	.43	–.04	–.06	–.29	**.23**	**.64**	**.64**
Hireability	.43	.43	.44	–.01	–.11	–.37	**.26**	1	**.7**

The correlational accuracy (Spearman ρ), effect-size difference (Cohen's d), and AI measurements of bias and fairness computed in our Spearman's "Women–Men" column shows the difference in the accuracy AVI case study across genders. Cohen's d "True/Pred" column shows the difference in effect size between women and men in the true versus predicted construct labels. The AI ratios are computed separately on the true and predicted labels. The **bold** Spearman and Cohen's d values denote small effect sizes ($|\rho| > 0.1$, $|d| > 0.2$), which should arouse suspicion and warrant further investigation. The **bold** adverse impact ratios fall below the 80% threshold and would be prima facie evidence of adverse impact by the "four-fifths" rule. True=ground truth; Pred=ML model predictions.

Louis Hickman (louishickman@gmail.com) received his Ph.D. degree in industrial-organizational psychology and his M.S. degree in computer and information technology specializing in natural language processing from Purdue University. He is with the Wharton School of Business, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA. His research focuses on the use of machine learning for personnel assessment, and his dissertation, "Algorithmic Ability Prediction in Video Interviews," received an SHRM Foundation Dissertation Grant.

Shree Krishna Subburaj (shree.subburaj@colorado.edu) received his M.S. degree in computer science from the University of Colorado Boulder. He is with the Institute of Cognitive Science, the University of Colorado Boulder, Boulder, Colorado, 80309, USA. His research efforts and interests span natural language and human signal processing for understanding multiadic human communication and experiences.

Louis Tay (stay@purdue.edu) is with the College of Health and Human Sciences, Purdue University, West Lafayette, Indiana, 47907, USA. He is the William C. Byham Chair of I-O Psychology at Purdue University and has published more than 100 peer-reviewed papers with numerous articles on measurement bias and Big Data. He is a coeditor of the volume *Big Data in Psychological Research* (Woo, Tay, & Proctor, 2020) and is an associate editor of *Organizational Research Methods*.

Sang Eun Woo (sewoo@purdue.edu) is an associate professor of I-O Psychology at Purdue University. She recently coedited the volume *Big Data in Psychological Research* (Woo, Tay, & Proctor, 2020). Her research interests include innovating methods of measurement and data analysis for studying important psychological phenomena in the workplace. She is an associate editor of *International Journal of Testing* and the 2021 chair for the American Psychological Association's Committee on Psychological Tests and Assessment.

Sidney K. D'Mello (sidney.dmello@gmail.com) is an associate professor with a joint appointment in the Institute of Cognitive Science and Department of Computer Science at the University of Colorado Boulder, Boulder, Colorado, 80309, USA. His research focuses on applying multimodal machine learning to investigate the interplay between the cognitive and affective states of individuals and teams engaged in real-world activities. He has coedited seven books and published almost 300 journal papers, book chapters, and conferences proceedings. He is an associate editor of *Discourse Processes* and *PloS ONE*.

References

[1] L. Hickman, N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo, "Automated video interview personality assessments: Reliability, validity, and generalizability investigations," *J. Appl. Psychol.*, early access, June 10, 2021. doi: 10.1037/apl0000695.

[2] S. Muralidhar, L. S. Nguyen, D. Frauendorfer, J.-M. Odobez, M. Schmid Mast, and D. Gatica-Perez, "Training on the job: Behavioral analysis of job interviews in hospitality," in *Proc. 18th ACM Int. Conf. Multimodal Interaction*, 2016, pp. 84–91. doi: 10.1145/2993148.2993191.

[3] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proc. 2021 Int. Conf. Multimodal Interaction*, to be published.

[4] American Educational Research Association, American Psychological Association, National Council on Measurement in Education et al., *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, D.C., 2014.

[5] A. Baird and B. Schuller, "Considerations for a more ethical approach to data in AI: On data representation and infrastructure," *Front. Big Data*, vol. 3, p. 25, Sept. 2020. doi: 10.3389/fdata.2020.00025.

[6] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Proc. European Conf. Comput. Vision*, 2020, pp. 506–523.

[7] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *Proc. Int. Conf. Multimodal Interaction*, 2020, pp. 361–369.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[9] "Title VII of the Civil Rights Act of 1964," U.S. Congress, Washington, D.C. Accessed: Aug. 31, 2021. [Online]. Available: https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021. doi: 10.1145/3457607.

[11] B. Hutchinson and M. Mitchell, "50 years of test (un)fairness: Lessons for machine learning," in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 49–58.

[12] R. Cropanzano, D. E. Rupp, C. J. Mohler, and M. Schminke, "Three roads to organizational justice," *Res. Personnel Hum. Resour. Manage.*, vol. 20, pp. 1–123, Dec. 2001.

[13] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Conf. Innovations Theoretical Comput. Sci. Conf. (ITCS)*, 2017, pp. 43:1–43:23.

[14] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[15] P. G. Devine, E. A. Plant, D. M. Amodio, E. Harmon-Jones, and S. L. Vance, "The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice," *J. Personality Soc. Psychol.*, vol. 82, no. 5, p. 835, 2002. doi: 10.1037/0022-3514.82.5.835.

[16] N. Tippins, P. R. Sackett, and F. Oswald, "Principles for the validation and use of personnel selection procedures," *Ind. Org. Psychol.*, vol. 11, no. S1, pp. 1–97, 2018.

[17] M. A. Campion, D. K. Palmer, and J. E. Campion, "A review of structure in the selection interview," *Personnel Psychol.*, vol. 50, no. 3, pp. 655–702, 1997. doi: 10.1111/j.1744-6570.1997.tb00709.x.

[18] D. K. Berlo, *The Process of Communication: An Introduction to Theory and Practice*. New York: Holt, Rinehart and Winston, 1965.

[19] M. Mehu and K. R. Scherer, "A psycho-ethological approach to social signal processing," *Cognit. Process.*, vol. 13, no. S2, pp. 397–414, 2012. doi: 10.1007/s10339-012-0435-2.

[20] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annu. Rev. Statist. Its Appl.*, vol. 8, no. 1, pp. 141–163, 2021. doi: 10.1146/annurev-statistics-042720-125902.

[21] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. IEEE/ACM Int. Workshop on Softw. Fairness (Fairware)*, 2018, pp. 1–7.

[22] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic fairness," in *Amer. Econ. Assoc. (AEA) Papers Proc.*, 2018, vol. 108, pp. 22–27. doi: 10.1257/pandp.20181018.

[23] Y. J. Weisberg, C. G. DeYoung, and J. B. Hirsh, "Gender differences in personality across the ten aspects of the big five," *Front. Psychol.*, 2011, vol. 2, p. 178. doi: 10.3389/fpsyg.2011.00178.

[24] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 469–481.

[25] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language," *J. Personality Soc. Psychol.*, vol. 108, no. 6, p. 934, 2015. doi: 10.1037/pspp0000020.

[26] C. Stachl, F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, and M. Bühner, "Personality research and assessment in the era of machine learning," *European J. Personality*, vol. 34, no. 5, pp. 613–631, 2020. doi: 10.1002/per.2257.

[27] Q. Song, S. Wee, and D. A. Newman, "Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices," *J. Appl. Psychol.*, vol. 102, no. 12, p. 1636, 2017. doi: 10.1037/apl0000240.

SP