American Psychologist

ISSN: 0003-066X

2023, Vol. 78, No. 1, 36–49
https://doi.org/10.1037/amp0000972

Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models

Richard N. Landers[1] and Tara S. Behrend[2]
[1] Department of Psychology, University of Minnesota, Twin Cities
[2] Department of Psychological Sciences, Purdue University

Researchers, governments, ethics watchdogs, and the public are increasingly voicing concerns about unfairness and bias in artificial intelligence (AI)-based decision tools. Psychology's more-than-a-century of research on the measurement of psychological traits and the prediction of human behavior can benefit such conversations, yet psychological researchers often find themselves excluded due to mismatches in terminology, values, and goals across disciplines. In the present paper, we begin to build a shared interdisciplinary understanding of AI fairness and bias by first presenting three major lenses, which vary in focus and prototypicality by discipline, from which to consider relevant issues: (a) individual attitudes, (b) legality, ethicality, and morality, and (c) embedded meanings within technical domains. Using these lenses, we next present *psychological audits* as a standardized approach for evaluating the fairness and bias of AI systems that make predictions about humans across disciplinary perspectives. We present 12 crucial components to audits across three categories: (a) components related to AI models in terms of their source data, design, development, features, processes, and outputs, (b) components related to how information about models and their applications are presented, discussed, and understood from the perspectives of those employing the algorithm, those affected by decisions made using its predictions, and third-party observers, and (c) meta-components that must be considered across all other auditing components, including cultural context, respect for persons, and the integrity of individual research designs used to support all model developer claims.

Public Significance Statement
Although artificial intelligence (AI) is now being used to make decisions about people's employment, education, healthcare, and experiences with law enforcement, external evaluators do not often agree on what is necessary to show that an AI is "unbiased" or "fair." This is in part because "bias" and "fairness" mean different things to different people. We created a framework for auditing that respects these differences in pursuit of better, fairer AI.

Keywords: audit, bias, psychology, machine learning, artificial intelligence

This article was published Online First February 14, 2022.
Richard N. Landers https://orcid.org/0000-0001-5611-2923
Tara S. Behrend https://orcid.org/0000-0002-7943-5298
The authors contributed equally to the work, and the order of authors is arbitrary. Andrew Schwartz and Andrew Smart contributed useful feedback on a draft of this article.

Richard N. Landers played an equal role in conceptualization, investigation, project administration, writing of original draft and writing of review and editing. Tara S. Behrend played an equal role in conceptualization, investigation, project administration, writing of original draft and writing of review and editing.

Correspondence concerning this article should be addressed to Richard N. Landers, Department of Psychology, University of Minnesota, Twin Cities, N-218, 75 E River Road, Minneapolis, MN 55455, United States. Email: rlanders@umn.edu

The use of high-complexity predictive models,[1] commonly referred to as artificial intelligence (AI), is rapidly expanding. Much like traditional statistical approaches,

[1] We use the term *predictive models* to refer to any statistical or machine learning model that can be used to make predictions about individual cases. This most centrally includes "supervised" regression and classification models, like ordinary least squares regression and support vector machines, which model sample criterion data from sample predictor data to make predictions about out-of-sample criteria using out-of-sample predictor data. However, it also includes "unsupervised" models, like principle components analysis, k-means clustering, and some approaches to neural network modeling, which model sample data without a criterion to make predictions about likely latent group membership for out-of-sample cases or variables.

Richard N. Landers

these models are being used to identify patterns in existing data to predict people's futures, to make decisions about them. The growing capabilities and popularity of AI has increased concerns about how exactly such predictions are generated and whether the use of these predictions has unintended consequences. These concerns are exacerbated by the fact that as the complexity of the models increases, transparency typically decreases, and it can be difficult to determine exactly how some models generate their predictions. There can be hundreds or thousands of variables in a modern AI model, plus potentially interactions between all these variables, making it difficult to identify the specific variables driving a particular recommendation or prediction. Although such predictive AI models exist in many domains of life, from GPS navigation to targeted marketing, we focus here on high-stakes decision making about individuals, a domain that is the focus of many AI-related complaints and fears. One example of a high-stakes decision-making context in the domain of industrial-organizational psychology is the recommendation to hire or not hire a job candidate (Gonzalez et al., 2019), or to admit or not admit a college applicant (Marcinkowski et al., 2020). Another example from forensic and criminal psychology is the use of predictive models to allocate policing resources to some neighborhoods and not others (Berk, 2021). A third example from clinical psychology in the domain of health interventions is the mining of text or voice data to identify people who are at risk of or experiencing mental health issues and automatic referral to human counselors (Graham et al., 2019). The use of AI to make high-stakes decisions like these is quite broad, touching most if not all domains of psychology across a wide spectrum of human activity.

Given the severe consequences of mistakes in high-stakes contexts like these, auditing has been proposed by researchers (e.g., Raji et al., 2020), by government (e.g., Vanian, 2021), by ethics watchdogs (e.g., Binns & Gallo, 2020) and in myriad op-ed columns (e.g., Kobielus, 2021) as a method to verify that AI-driven predictions are fair, unbiased, and valid. A well-conducted interdisciplinary audit, with a strong foundation in both psychometric validation principles from psychology and model development principles from statistics and computer science, whether internally adopted as a standard test development practice or externally imposed by a regulatory authority, can identify adverse effects from the application of a model on marginalized groups or individuals. Auditors can recommend corrective actions to remove or minimize those adverse effects. Audit results can also be used to assist decision makers in maximizing the value of a model by identifying and repairing aspects of the model that lead to undesirable consequences. Although auditing often occurs after an AI-scored test has been developed and is being used, careful adherence to core auditing components at all stages of assessment development and complete documentation of these efforts can diminish or even preclude the need for post hoc auditing. Thus, the application of high-quality auditing processes at all stages of a test's development and deployment can both verify developer claims and improve algorithmic success, increasing public trust in the particular AI model investigated, the decision-making entity responsible for it, and AI technology in general.

Challenges in Designing Audits

Despite its intuitive appeal, applying auditing standards is fraught with challenges. Most critically, conducting an audit to determine whether an AI model is fair is complicated by the fact that the words *fairness* and *bias* have very different meanings for different audiences, especially across disciplines and areas of expertise (Mulligan et al., 2019). From the perspective of social science, the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014; hereafter "the *Standards*") provide critical guidance on many of the measurement issues central to auditing, but because of their authorship define and consider fairness and bias relatively narrowly. Similarly limited but from the perspective of computer science, the Organisation for Economic Co-operation and Development's (2019) *Principles for AI* define AI fairness in terms of transparency regarding what is being measured and what inputs are used in algorithms, respect for individuals and diversity, and ensuring that AI is used to further human rights and human dignity. Other groups more narrowly focus on adverse effects on individuals when discussing fairness, for example in identifying flaws in the representativeness of the data used to build the model that led it

Tara S. Behrend

to disadvantage members of underrepresented groups (Miller, 2021). A wide range of citizen groups have emerged that focus on potential invasions of privacy; these invasions too are labeled as unfair. Large corporations such as Facebook and Google have in-house ethics teams serving as internal watchdogs; these groups have their own definitions of fairness and bias (Schroepfer, 2021). This complex landscape of actors, goals, and values can make it difficult to communicate and difficult for psychologists to demonstrate their value and in some cases even meaningfully participate in discussions about fairness. Furthermore, this landscape makes conducting an audit challenging, because the auditor must take a particular stance about what fairness means, a stance that may not be universally accepted.

In the present paper, we explore the meaning of bias and fairness in AI through three major lenses and outline key considerations for audits of AI models. In doing so, we provide a framework for psychologists and others who seek to provide meaningful auditing services to users and builders of systems that use AIs to make predictions about people. AI development can and should benefit from the many robust frameworks for fairness, bias, and validity created and endorsed by psychological science, as many (although not all) of the issues crucial to quality AI-based modeling of human characteristics are already thoroughly explored in profession-wide consensus documents like the *Standards*. High-stakes assessment contexts, in particular, are fraught with ethical and social challenges that psychology is well-poised to inform. These frameworks can be used to minimize harm to groups that might be adversely affected by unfair AI. Furthermore, public trust in audited AI may be increased by applying rigorous psychological science and its associated ethical standards at each stage of the auditing process.

Focal Example

To illustrate key concepts, we focus upon a complex illustrative example. Consider the use of a machine learning algorithm to score a virtual asynchronous job interview (e.g., Brenner et al., 2016). In this interview, a job candidate responds to a prompt on a website like "Describe a time you resolved a workplace conflict" by recording and submitting a video of themselves, which is then analyzed using AI. The scoring algorithm includes thousands of variables relating to word choice (e.g., shorter words, more words, more positively toned words, grammar), answer content (e.g., intended to provide evidence of particular knowledge or skill), voice patterns (e.g., tone, inflection, pitch, speed, volume, pauses), and visual data (e.g., facial expressions, facial features, eye contact). Due to the complexity of the model, it attends to other variables, such as the presence of wrinkles (possibly indicating age); variation in skin tone (possibly indicating race or ethnicity), and color patterns attributable to makeup (possibly indicating gender). Although the designer did not create the algorithm to assess these characteristics, and their inclusion is undesirable, the algorithm nevertheless does so, and the designer does not realize this. Like most models in use this way, it was built using a sample of existing job candidates for whom performance data were available. The model identifies patterns that are associated with better job performance and looks for similar patterns in new candidates. After deployment, it is discovered that the scoring algorithm recommends fewer women for promotion than men. Yet unlike traditional psychological statistical models, because the machine learning algorithm is so enormously complex, it is not evident to its developers why this is occurring nor are initial model diagnostics revealing clear answers. Such a problem might have been caused by the nature of the input data, the various data processing steps between those data and the model being run, the model's design and execution, or any combination of these factors or numerous others. In short, the developer's claim that this algorithm predicts future job performance equally well for all groups and using appropriate modeling techniques is questionable and should be audited.

This example is particularly useful because it describes an AI application that is realistic, reflects numerous decisions commonly made by test scoring algorithm developers, and would have effects of interest to both policy makers and the general public. It is also useful because a rigorous and routine internal auditing process would have prevented this algorithm from having been deployed before addressing the issues that emerged later. In the sections that follow, we will parse potential sources of bias or unfairness in this situation to better understand how an audit can identify and remedy such issues as we explore the broader conceptual issues in auditing.

While doing so, we explicitly consider the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 2018; hereinafter "the *Principles*") as representing the most current and authoritative set of professional standards for psychological measurement in organizational contexts and thus the most reasonable starting point for defining psychological auditing in that context.

Defining Fairness and Bias

Whether used by laypersons or experts, the word "fairness" can refer to a wide range of different feelings, actions, and outcomes (Hutchinson & Mitchell, 2019). Further complicating a meaningful discussion is that critics of AI who decry something as unfair may be referring to what another person would consider to be bias; yet the word "bias" is similarly laden with conceptual complexity and confusion. Among this variation, we have identified three major categories of conceptualizations of fairness and bias. First, fairness and bias may be understood through the lens of individual attitudes. For instance, a person may label an action unfair if they perceive that a decision created by an AI has been made in a way that violates their personal principles of fairness. Second, fairness and bias may be understood through the lens of legality, ethicality, and morality, closely related yet distinct perspectives that all rely less on any one individual's judgment and instead evaluate decisions from more broadly developed rules and conceptual frameworks formalized by existing stakeholders. Third, fairness and bias may be understood from various domain-embedded technical perspectives, in which different research communities, especially psychometric testing and machine learning, have their own technical definitions of each term. Because concerns from any or all these perspectives may motivate a fairness and bias audit, and to better bridge these perspectives when conducting such an audit, we briefly explore them all.

Lens 1: Individual Attitudes

The lens of individual attitudes is the most invoked in public discourse. We can understand this through the scholarly lens of justice theory, a tripartite perception framework consisting of distributive, procedural, and interactional justice perceptions (Greenberg, 1990), which combines cognitive, perceptive, and emotional aspects. From the perspective of psychology, many complaints about AI fairness implicitly argue from a distributive justice standpoint. Distributive justice refers to perceived fairness of outcomes, such as employment opportunities, clinical diagnoses, or college admission. The fairness of these outcomes is judged according to personal rules of equality, need, or equity, which are themselves informed by cultural and social values, as well as referent others. When applying equality rules, people expect outcomes to be distributed equally to all. When applying need rules, people expect outcomes to be given to those who need them most. When applying equity rules, people expect that outcomes are proportional to inputs, such that those who contribute the most are rewarded with the best outcomes. If applying different rules, two people may judge the fairness of the same decision quite differently, and one person may even apply different rules to slightly different situations or over time (Holtz & Harold, 2009). In our focal example, distributive justice perceptions describe whether a person believes the algorithm's final prediction or its apparent effect, in this case to hire or not to hire, is fair (cf. Acikgoz et al., 2020).

Although distributive justice is highly salient and observable, it is not the only way to define attitudinal fairness. Procedural justice, the second component of justice, refers to the perceived fairness of rules and procedures that are used to determine outcomes. For example, if college admissions decisions are based on a test that is faulty in some way, and there is no way to register a complaint to fix the error, this would be a procedural injustice. Procedural injustice is not dependent upon distributive justice; for example, a person might dislike their own negative outcome but still believe that the process itself was fairly conducted. In the context of our focal example, we identified five of Ford et al.'s (2009) procedural justice rules as likely implicit criteria by which procedural justice would be judged. First, an AI-based decision may create perceived violations of opportunity-to-perform rules, likely to occur when a job applicant does not believe they have been given a fair chance to demonstrate their value. Second, an AI-based decision may violate job-relatedness rules if the method by which algorithms come to decisions lack face validity; for example, people likely differ in the degree to which they believe facial expressions during a job interview are relevant to the prediction of future job performance, and whether evidence of relevance is required to justify its inclusion in the model (Tippins et al., 2021). Third, when algorithmic decisions cannot be appealed, this may violate rules about reconsideration opportunities more readily than in traditional hiring scenarios (Langer & Landers, 2021), and fourth, violations of two-way communication rules are likely common when AI is used to replace face-to-face interaction with human decision-makers. Fifth, some people may simply believe that AI-based decision-making is morally inappropriate, a violation of propriety rules. People vary greatly in the degree to which these different rules influence their perceptions and how various rule violations combine to create specific reactions toward particular systems (Brockner et al., 2007). In our focal example, procedural justice perceptions describe whether a person believes the way the algorithm generated its

prediction is fair. This would likely be based upon a combination of the consumption of information the algorithm developers provided about their algorithm, reflections upon the system used to collect their data, and prior attitudes about that system or similar systems.

The third component of justice, interactional justice, refers to perceived fairness of the interaction between decision makers and those affected by their decisions. It is sometimes split into interpersonal justice, which occurs when people feel they have been treated with respect and dignity, and informational justice, which occurs when people feel they have been given appropriate information about the decision and the process used to arrive at it. These two components are also quite salient in the domain of AI-based decision making. Informational justice is likely most directly affected by transparency in terms of both what is being assessed and what is being done with that information. Interpersonal justice is likely driven by the presentation strategy of that information in terms of how details about the AI are explained and how timely those explanations are. In our focal example, interactional justice perceptions describe whether a person believes they were given adequate information about the algorithm and treated respectfully, such as by viewing an informational video before data collection explaining how decisions had been made to ensure safety, privacy, and fairness.

Lens 2: Legality, Ethicality, and Morality

Beyond individual perceptions, fairness and bias may also be understood through the lens of legality, ethicality, and morality. Unlike the lens of perception, this lens considers fairness from the perspective of culture and society, both locally and more broadly. Here, a fair AI system is one that aligns with shared human values and supports human flourishing. A fair system conforms with both relevant laws and established professional guidelines. Morality ideals, which form the foundation of fairness through this lens, are shaped by history and culture. They are governed by a sense of responsibility to others—by a sense of *care* (Gilligan, 1987/1995). The ethics of care are not in conflict with justice perceptions; instead, they are reciprocal. Bodies viewing fairness through this lens generally define bias in a way that reflects this shared sense of care for others.

As a key example, the five core principles of American Psychological Association's (APA; 2017) code of ethics serves as instructions for how psychologists should care for the people they interact with in the course of their work and reflect this sense of shared cultural values. First, the principle of beneficence and nonmaleficence directs psychologists to strive to benefit others and do no harm to them. Second, the principle of fidelity and responsibility directs psychologists to establish relationships based on trust. Third, the principle of integrity directs psychologists to be honest and truthful. Fourth, the principle of respect for people's rights and dignity directs psychologists to treat people equitably regardless of their personal characteristics or group membership. Fifth, the principle of justice directs psychologists to address and minimize biases in their own thinking and treatment of others. Although *bias* is not defined explicitly in the code, it appears to refer to a goal of impartiality, lacking individual prejudices and cognitive biases (cf. Boudana, 2016). Any psychologist working on the system in our focal example would be beholden to these principles when making decisions about algorithm design.

Ethical codes unique to AI have also emerged, articulating similar core principles, as have local and state laws intending to shape shared expectations about the ethical use of AI. These codes and laws tend to use the term *bias* in similar ways. The most prominent AI ethical codes, the OECD's (2019) *Principles on Artificial Intelligence* and The Public Voice's (2018) *Universal Guidelines for Artificial Intelligence* (UGAI), make references to fair and unbiased decisions but do not define these terms precisely. These codes reflect an ethics of care for others, such as in their inclusion of transparency and public safety requirements. UGAI also explicitly mentions reliability, validity, and data quality as key aspects of AI systems and uses the terms *bias*, *discrimination*, and *unfairness* essentially interchangeably. For example, the use of facial recognition algorithms as in our focal example has been widely criticized as its applications have expanded to include policing and other high-stakes contexts. The likelihood of a false positive (e.g., wrongly identifying a focal photo as matching that of a suspect when used to make positive identifications of crime suspects, leading to false accusation of a crime) can be over 100 times higher for Black and Asian faces as for White faces for some algorithms (McLaughlin & Castro, 2020) due in part to poor representation of darker-skinned people in source datasets (Buolamwini & Gebru, 2018). Here, the false positive rate is the focus of criticism, presented as an ethics violation.

In contrast, laws that concern discrimination in housing, hiring, and college admissions tend to have precise technical definitions of bias based on statistical inference, in many cases built upon decades of precedent and case law. One such legal framework relevant to our focal example is that of employment discrimination law. Legally establishing the bias of a test for employment generally relies on a statistical concept called *differential prediction* (Berry, 2015), which involves comparisons of regression lines between groups within legally defined classes, such as by sex, race, or national origin. Analysis of differential prediction is often used to help identify the source and justifiability of *disparate impact* (also called *adverse impact*, defined as differential selection rates across group membership within a class). In contrast, intentional discrimination, called *differential treatment*, refers to explicitly treating members of classes differently due to their class membership. In our focal example,

disparate impact would be observed if the algorithm predicted greater job performance for men than for women. Disparate treatment would result from building processes within the algorithm to treat groups differently, such as by awarding bonus points to members of underrepresented groups or modeling class membership explicitly, as a predictor variable.

The subtle differences between decades-old concepts like these are still often muddled even in modern policy and law. For example, the influential policy organization, the Electronic Privacy Information Center (n.d.), has combined concepts in their advocacy, writing, "AI systems have been shown to exacerbate bias and disparate impacts based on gender, age, race and other characteristics in hiring, housing, criminal justice, surveillance, and other contexts." Another example is a New York City law that requires an annual "bias audit" to evaluate any system that "automatically filters candidates or prospective candidates for hire or for any term, condition or privilege of employment in a way that establishes a preferred candidate or candidates" yet does not define what it means by "bias" (Sale of Automated Employment Decision Tools, 2020). In this way, many current bills, laws, and policies appear to leave the concept of *bias* vague, possibly so that they remain applicable as professional standards and ethical codes change. In practice, this approach makes the ethics of AI technocratic, leaving the interpretation of "bias" at the discretion of those adopting the third lens, that of technical definitions embedded within the various disciplines and professional organizations monitoring the use of AI.

Lens 3: Technical Domain Embedded Meanings

The term bias has various technical definitions depending on the specific research domain in which it is used. Almost all these definitions are based upon a shared interpretation of error, which broadly refers to inaccuracy in the estimation of some population value using sample data. Error is in turn commonly split into random and systematic types. The presence of random error is unavoidable in virtually all estimation techniques; most statistical procedures, including machine learning, are designed to minimize random errors but accept that random errors cannot be eliminated. Although this may result in inaccurate predictions for individual cases, random errors when averaged will theoretically equal zero, such that predictions for some individuals are too high whereas others are too low, and these are roughly symmetrically distributed. In contrast, systematic error refers to nonrandom inaccuracies.

To illustrate these concepts, consider again our virtual interview machine learning algorithm focal example. The prediction of future job performance by scoring interviews, whether by human or by algorithm, will never achieve 100% accuracy (e.g., $R^2 = 1.0$); this is not realistically achievable, due to both imperfect measurement of job performance and because job performance is multidetermined by many factors unrelated to available signals in job interviews. Some signals will be more relevant for some people's job performance and less relevant for that of others. When these inaccuracies vary only by person and do not favor one group of people over another, they are random. However, if this inaccuracy correlates with group membership, such as with race, such errors are systematic (Köchling & Wehner, 2020). The question of whether such systematic errors as in our focal example are evidence of bias depends on the term's local definition within a given research field and context.

Contrasting Statistics, Machine Learning, and Psychometrics Perspectives

In statistics, the term bias is often used to refer to statistical bias, or systematic error in sample representation in comparison to population composition. One type of statistical bias, sampling bias, is particularly common in the context of algorithmic audits. Consider a final performance estimate (e.g., $R^2 = .5$) of a machine learning algorithm used to score virtual interviews. The generalizability of the R^2 is threatened if the sample used to build the model differs from the population nonrandomly. For example, researchers hoping to build an algorithm that provides generalizable prediction of a racially and ethnically diverse population without sampling bias should sample from that population at random. To the extent that they are unable to do so, such as in the common case of undersampling racial and ethnic minorities, the resulting model may show the effects of sampling bias, more heavily weighting data and variable interrelationships associated with those individuals most frequently represented. In the case of AI, sampling biases are often observed in the computer vision literature, as in the earlier example describing how facial recognition sample databases tend to contain more White people than the population to which they need to generalize (Buolamwini & Gebru, 2018), a problem particularly salient in the context of predictive policing (Berk, 2021). Considered this way, bias is a problematic consequence of improper sampling. In our focal example, this particular problem might be addressable by either collecting a more diverse database or by oversampling data from people who are not White, although neither is guaranteed to be successful in practice.

In machine learning, the term bias often appears in discussions of the *bias-variance tradeoff*, a central consideration in the development of machine learning models (Belkin et al., 2019). Traditional frequentist statistical procedures, especially those seen in mainstream psychology, fall on the "low bias-high variance" side of this tradeoff, which reflects the prioritization of calculating so-called "unbiased estimates" of effects. In short, most common psychological statistical procedures, like *t*-test, analysis of variance, and

ordinary least squares regression, were designed to minimize random error given the sample in which modeling is performed, maximizing predictive accuracy within the sample so as to give clear and unambiguous interpretations of individual coefficients. However, when applying such a model to new data collected from the original population, this lack of bias within the sample is associated with increased variance in out-of-sample predictive accuracy. In short, placing "unbiased estimates" above all other priorities can result in less stable out-of-sample predictive performance, a problem called overfitting. Because the use of models to predict values out-of-sample both accurately and consistently is often the primary goal of machine learning models, modeling techniques were developed that introduce bias in specific controlled ways in exchange for reduced out-of-sample variance (Yarkoni & Westfall, 2017). For example, one of the clearest signs of overfitting in an ordinary least squares regression model is the presence of large estimated regression weights; in response, ridge regression, lasso regression, and their machine learning integration elastic net, were developed to penalize large regression weights. Although individual model parameters are no longer interpretable as unbiased as a result of this penalization, the predictive accuracy of the model out-of-sample is improved. In this way, bias can be a positive and desirable aspect of a well-engineered model when used to improve other model characteristics. In our focal example, procedures like this are necessary to address the extraordinarily high complexity of a dataset containing language, audio, and video, and therefore potentially tens of millions of variables; creating bias purposefully, and in the process losing the clear interpretability of individual coefficients associated with individual points of color and light at a particular second in a video, is generally viewed as a worthwhile exchange.

In psychometrics, the term bias is often used to describe differences in measurement characteristics between identified groups. There are a variety of technical approaches to identifying bias within psychometrics, but one of the most common is as measurement invariance, which refers to differences across groups in the latent characteristics of the measure within a confirmatory factor analytic framework or of item parameters within an item response theory framework (Vandenberg & Lance, 2000). For example, if certain parameters are systematically larger or smaller for men than for women, it suggests that the use of that measure to represent both men and women will result in biased estimates (i.e., estimates containing systematic error) for one or both groups. Psychometric bias may or may not be problematic. For example, if a psychometric measure is designed to assess educational attainment, bias as reflected by latent mean differences in educational attainment between racial groups would be expected, caused by associated systemic differences in educational opportunities across those groups (McDaniel et al., 2011). Thus, measurement bias is generally not considered problematic by test developers if it reflects bias in the construct being measured. Such a test might be considered biased but fair (Meade & Tonidandel, 2010).

A similar approach to bias is used across prediction contexts, including both ordinary least squares regression and more complex supervised machine learning models. This is particularly relevant in the context of employee hiring, as in our focal example, due to the increased relevance of the differential prediction model described earlier. Noting the definitional challenges of the word "bias," several in the psychometrics community have argued that differential prediction is not a sufficient condition for bias, a term they contend has an extreme connotation, arguing that a test can demonstrate differential prediction without having problematic measurement qualities. Thus, differences in regression lines between groups in terms of intercepts, slopes, or standard errors are often used as evidence of differential prediction rather than of bias (Stark et al., 2004). For example, in the context of our focal example, consider a managerial job for which the use of the term "six sigma" is associated with job performance. Because younger applicants are less likely to have been exposed to six sigma management principles, the use of this scoring model might create differential prediction by applicant age. Yet because age is associated with managerial experience, if managerial experience is related to job performance, differential prediction still may be considered fair (Meade et al., 2009).

Designing an Effective Psychological Audit

Designing a useful and accurate audit is thus quite challenging given the many different conceptualizations of fairness, many of which are based upon different fundamental assumptions deeply embedded within disciplines. An audit must therefore define and defend its basis for conceptualizing fairness, and the best audits must cross disciplinary barriers when doing so. When designing an effective audit of a system based on AI, all assumptions and values must be questioned and considered openly. Given our expertise, we focus here on providing guidance to those reaching across disciplinary lines from psychology to other fields; however, we contend that the principles we introduce here have broad value regardless of one's initial assumptions.

Given this framing, we define a psychological audit as an impartially conducted conceptual and empirical evaluation of claims about psychological characteristics, including traits, attitudes, emotions, behaviors, and other quantities that are not directly observable, as measured or predicted by algorithms. The specific psychological characteristics targeted by such algorithms are determined by the developers of the algorithm, which may not correspond to the characteristics ultimately measured or predicted. Audits must be both conceptually based in psychometric theory and empirically tested, using data collected with a research design appropriate

to evaluate the claims being made and interpreted according to modern standards of test evaluation. Although our focal example concerns employment, audits can be used to evaluate any such claim about predictions produced by an algorithm.

Audits may be designed to evaluate claims of any type but are most valuable to assess claims of validity (i.e., is measurement or prediction of the targeted characteristics being done consistently and accurately?), utility (i.e., does the algorithm provide value from its implementation?), and lack of bias. The evaluation of validity claims is relatively more straightforward due to general agreement about the nature of validation testing within the psychometrics community and published standards for test validation procedures (e.g., Binning & Barrett, 1989; Messick, 1995). Similarly, although there is disagreement on how value is conceptualized even within disciplinary lenses, there are already several frameworks for estimating it in psychology and business, including existing research on multiattribute utility (Roth et al., 2002), return on investment (Kim & Ployhart, 2018), litigation liability (McPhail, 2005), and triple bottom line accounting (Norman & MacDonald, 2004). In medicine, as seen in the algorithmic prediction of healthcare intervention effectiveness, value is often conceptualized as lives saved or suffering reduced (Moskop & Iserson, 2007). Bias is the most complex, because claims can be made through any of the lenses of fairness and bias previously described. In fact, failing to articulate the precise standards of fairness and bias by which an audit is conducted, and therefore failure to articulate the value system used to develop the algorithm, can render the results of an audit uninterpretable across disciplinary lines.

Because of the ambiguity introduced by shifting or unclear bias lenses, evaluating the credibility of auditors is paramount to support an impartial evaluation of claims. In practice, auditors are identified in three ways: (a) internal auditors are employees of the company that developed the algorithm with expertise in psychometrics and algorithms, (b) external auditors are from consulting firms and hired by the company that developed the algorithm, and (c) independent auditors are brought in externally, often but not exclusively by a regulatory authority, to evaluate the algorithm. Although it is intuitive to believe these fall in a hierarchy of credibility, access to data and documentation can vary greatly even within auditor type, with different restrictions in nondisclosure agreements and thus freedom to explore deeper aspects of algorithmic systems. Thus, all auditors regardless of source must be evaluated independently according to their presentation of the measurement standards and definitions of bias that they apply in due course of auditing, and the terms of access and nondisclosure, which should specify what types of information can be withheld at the company's discretion, should themselves be disclosed. No auditor or audit is automatically credible; claims that an algorithmic system has been audited and is therefore credible should be viewed skeptically.

Psychological audits can be conducted for many purposes, often driven by the intended audience of the audit. It should not be assumed that audits are intended to "catch" wrongdoing; a particularly valuable purpose of auditing can be to provide guidance to the developers of the algorithms, both during and after their development, so that the algorithms can be improved (i.e., formative auditing). Normalizing routine internal or external auditing this way would be a great public good, increasing the probability that algorithmic systems in general are valid, valuable, and fair. Furthermore, when a particular algorithmic system has an outsized impact on society relative to other algorithms with similar goals, as is the case for the algorithms used in many large Silicon Valley technology companies, the social benefit from more publicly facing, transparent, open audits increases. The results of such audits should also be presented in multiple formats to meet the needs of all relevant audiences, such as by producing both a precise and comprehensive technical report for review by testing professionals and a layperson-friendly summary for whomever the audit's predictions will directly affect. In general, unless there is a compelling and transparently stated reason not to release the results of an audit when the results of that audit are in the public interest, the audit should be released; organizations may choose to do otherwise, but they risk harming the credibility of that audit, the company that designed the algorithms, and the auditors involved.

Major Components of an Audit

Although audits should be designed to evaluate specific claims, there are many components of algorithmic systems that can be examined in due course of such an evaluation, as shown in Table 1 in a likely but not necessarily chronological order of investigation. Importantly, this table is not intended to be exhaustive but rather to provide a framework for the most important issues to be addressed in most audits. Although versions of these components exist in existing validation frameworks from psychology, machine learning frameworks from computer science, and domain-specific frameworks within individual application areas, we contend our framework is the most comprehensive, interdisciplinary integration to date. If an algorithm developer makes a claim that cannot be fully evaluated within this framework, additional questions should be asked as relevant to modern professional standards. We have split these concerns into three major categories as shown in Table 1.

Components Relating to Models

The first component, input data, refers to the validity and generalizability of the initial dataset used to train a model relative to the population in which predictions will eventually

Table 1
Components of AI System to Be Audited

Component	Questions to ask	Applied to focal example
Components relating to models		
1. Input data	How were input data collected in terms of population and research design? How did these factors affect data quality?	Were the input data collected from job incumbents? How are incumbents different from applicants? Is there range restriction on variables of interest?
2. Model design	What drove initial model choices (e.g., criterion, predictor set, algorithm)? Were they informed by theory or empirically derived? If empirically derived, from what data (and of what quality were those data)?	How is performance defined, and how are we sure of its validity? Why are word choice, answer content, voice patterns, and visuals being included as predictors? Are these decisions theoretically supported?
3. Model development	Once the initial model was created, how was it refined? What approaches were taken, and what likely effect did these approaches have?	Is every decision about every model created during the development process fully documented? When were models discarded and why?
4. Model features	How were the raw input data engineered into model features? Was this process conceptually or empirically driven? What alternative feature engineering approaches were explored?	What specific natural language processing techniques were used? What evidence is there of quality speech-to-text conversion? What facial features emerged from analyzing video? What biases were explored from these engineering processes?
5. Model processes	How does the model use inputs to generate scores? How were alternative approaches explored and evaluated?	What stress tests were conducted, and what types of bias were investigated? Did these tests result in changes to the model, and if so, how and why?
6. Model outputs	How was the quality of predictions generated by the model evaluated, such as for psychometric reliability and validity? How was cross-validation conducted, and was it appropriate given claims about model generalizability?	Are scores consistent over time and upon multiple administrations (i.e., reliability evidence)? Is there evidence that they reflect the predicted constructs they claim to (i.e., validity evidence)? Do they show differences among classes of interest (e.g., race, gender, color, national origin, religion, disability, age) and combinations of classes?
Components relating to information and perceptions		
7. First-party interpretation	Does all messaging from the algorithm developer logically, honestly, and transparently follow from answers developed elsewhere in this audit?	Does the developer claim the model predicts job performance? What evidence in the audit forms the basis of this claim? Is anything exaggerated? Are important details left out?
8. Second-party effects	Who is directly affected by the use of the algorithm, and how have their outcomes and reactions been assessed? What is the relative impact of acting upon false positives versus false negatives on second parties?	How do nonselected applicants react to the news that the algorithm did not assign them a high enough score to be selected? What information is communicated to them, and how do they evaluate that information?
9. Third-party understanding	How have perceptions and evaluation by outside observers been assessed and incorporated? Have outside regulatory groups and community organizations been consulted?	How do experts in employment law view the documentation and performance of the algorithm? How does the public view this use of algorithms?
Meta-components		
10. Cultural context	Has the broader cultural context in which the algorithm will be used been considered? Have members of the community participated in the design of systems that will affect them?	Do power differentials exist between designers, employers, and job candidates? Have cultural assumptions been made? Will development decisions in one culture be applied to another, and if so, how has the development process been adjusted to prevent cross-cultural application challenges?
11. Respect	Is the algorithm being used in a way that conforms to generally accepted ethical standards, such as those in the *Standards*, the SIOP *Principles*, the OECD *Principles* and the UGAI?	What ethical standards do the developers claim to have followed in development? Is there evidence of decisions made following that ethical framework? What evidence is there that individual fairness has been a priority in development?
12. Research designs	How do the research designs (including sampling, experimental design, variable choices, analysis, and interpretation) of any studies conducted to support a claim affect the validity of conclusions?	For every claim that appears to be based upon empirical observation, does the study design support the claims made? Were all design decisions defensible from the perspective of modern methodological research? What impact might they have had on the validity of drawn conclusions?

Note. AI = artificial intelligence; SIOP = Society for Industrial and Organizational Psychology; OECD = Organisation for Economic Co-operation and Development; UGAI = Universal Guidelines for Artificial Intelligence.

be generated and used. Input data should be comprehensively evaluated in terms of the population they are drawn from, the sampling techniques used to draw them, and the adequacy of the data to conduct later analyses used to support claims. In our focal example, data were collected from a convenient source of existing job candidates, which by itself is no guarantee that those job candidates are representative of the future job candidates to which the model is intended

to be generalized. Thus, evaluating how the developers investigated this assumption and concluded it was not a problem is an obvious auditing goal that should emerge during investigation of this component. Importantly, this component also includes the use of later subset datasets for more specific modeling challenges; every training dataset should meet the same quality standard, embedded within a meaningful data engineering framework (Hutchinson et al., 2021). In our focal dataset, as described in the *Principles* (SIOP, 2018), population selection, the choice between concurrent and predictive validation study designs, and many other related issues influence the quality of conclusions that can be inferred from models based upon those data.

The second component, model design, refers to the initial far-reaching decisions about data selection and modeling that are difficult to test or difficult to change later in the process. The most common of such design decisions will be (a) the nature of the criterion modeled, (b) the set of predictors used in such modeling, and (c) the initial set of algorithms used for that modeling. For example, even if data are deemed an adequate representation of the population (addressing Component 1: Input Data), the choice of criterion measure and available predictor sets will drive many subsequent steps. This decision may be made conceptually, empirically, or heuristically. In our focal example, very few details are provided; we only know that "performance data were available." An audit must include evaluation of the quality of that performance data in terms of its construct validity as a valid representation of "performance," including a clear, unambiguous, and scientifically defensible definition of *performance*. Beyond this issue, the focal example states the use of word choice, answer content, voice patterns, and visual data but gives no rationale for that inclusion. Although such variables can be dropped in later analytic stages if they prove problematic for whatever reason, their initial inclusion and how they are handled in those later stages generally reflects decisions made at the design stage, increasing technical debt and the risk of sunk cost fallacies, both likely affecting decision-making. Such assumptions, processes, and outcomes must be audited.

The third component, model development, refers to the specific approach taken to create a machine learning model once initial model design is complete. It is when investigating this component that guidance in the *Principles* and *Standards* becomes quite thin, and machine learning guidelines become more useful. This component focuses upon the specific decisions made by model developers, often reflected in R or Python code or only in the minds of the developers, once a machine learning problem has been defined in terms of its inputs and outputs. Unlike the previous two components, the issues to be investigated in this component tend to be highly idiosyncratic to a particular model. A common issue in development is the choice between k-fold, holdout, and temporal sample cross-validation approaches. Frequently, because cross-validated model performance tends to be better than holdout sample performance which tends to be better than temporal validation performance (due to the decay of model performance over time), only k-fold performance will be reported, potentially reflecting hidden information or an incomplete model development process. In our focal example, all three validation approaches could have been used to compare the results; if they were not, the reasons for the omissions should be identified. Another frequent issue in model development is the reconceptualization of criteria; for example, a common issue in the creation of performance models like the one in our focal example is the relative weighting of task and contextual performance, which often has bias implications (Johnson, 2001). Specifically, by down-weighting task performance and upweighting contextual performance, adverse impact against underprivileged and underrepresented groups can sometimes be reduced; however, alternative weighting schemes have significant implications for validity (Sackett et al., 2008), so such decisions must be carefully articulated and defended. In computer science, this problem would be framed differently, as an issue with the selection of a *ground truth*, often without any explicit consideration of construct validity.

The fourth component, model features, refers to the engineering process used to create the final predictors used in modeling. This is another area where resources from machine learning will be more helpful than existing psychological guidance. For example, in our focal example, if bag-of-words modeling was used to predict performance from word choice, an approach in which word counts are calculated and converted into sets of variables called tokens, hundreds of decisions need to be made about the nature of language and its meaningfulness to the individual prediction problem being approached. In its simplest form, token data are frequencies created to reflect actual word use such that a value reflects the number of times a person uses a particular word. However, tokens might alternatively be created based upon the topic they most closely reflect. A model created on topics instead of tokens has the potential to reduce bias by minimizing the influence of dialect, yet this choice can have validity impacts that must be evaluated. An overarching concern in our focal example is also the use of machine learning-based speech-to-text converters to convert job candidate audio into text, which is often done with third-party machine learning models, among which "state-of-the-art" performance is a constantly moving target with important implications for psychometric reliability and validity. All steps occurring between "raw input data" to "engineered features," including but not limited to the issues highlighted here, must be fully investigated in a complete audit.

The fifth component, model processes, refers to the specific estimation process by which predicted scores are generated in the final model. For example, there are numerous specific machine learning algorithms that can be used to create

predicted scores for cases, and the propriety of these algorithms varies by the specific nature of the prediction problem. Each algorithm brings unique assumptions and best-use cases that must be evaluated in an audit. Most commonly, this is done via a process called stress testing, in which simulated or prototypical data not contained within the original training set are fed into the model and the predicted scores examined to assess their validity. Such tests can be used to better understand the conditions under which the model is generating predictions as its developers believe and when it is not. In our focal example, mock interviews by actors could be recorded and fed into the model to observe the effects. Stress tests should ideally be conducted across all classes of concern, including by race, gender, skin color, religion, national origin, disability, and age, especially when such information is engineered into model variables. Because video content is being converted into model features, the depth of stress testing should be much higher.

The sixth component, model outputs, refers to the validity of the final predicted scores generated by the model. Even with high quality training data, development processes, features, and processes, the final predictions may still be "unfair" depending upon one's reference frame for both validity and fairness. Both machine learning literatures (see Kleinberg et al., 2017) and psychological literatures provide guidance on the evaluation of outcome fairness. In our focal example, which occurs in the context of employment, an obvious audit evaluation step would be to assess compliance against current professional standards for employment tests, such as outlined in the *Principles* and the *Standards*, both of which focus heavily on output quality and set clear standards for high-quality measurement. Most critically, predictions generated by models in this context are always claimed to represent a construct, such as future job performance or turnover risk. As such, psychological standards for the predicted variable's construct validity apply, and compliance must be audited.

Components Relating to Information and Perceptions

The seventh, eighth, and ninth components refer to how information about the model is communicated and how that information is received. This flow of information can be considered from first-party, second-party, and third-party perspectives (Langer & Landers, 2021). Evaluation of first-party interpretation is relatively straightforward, requiring that auditors assess if the press releases, technical reports, and other information shared by the developers accurately reflect the details uncovered in the audit. Evaluation of second-party effects requires that auditors identify who is most directly affected by use of the predicted scores generated by the model; in this case, job candidates are the second party. Evaluation of third-party understanding requires auditors to assess if not only the information shared by first parties is accurate but also those absorbing, reacting to, and evaluating that information are doing so accurately and completely.

In our focal example, auditors should at a minimum assess if technical reports shared by the developers contain accurate information given the results of the other components of the audit (first party interpretation), determine how job candidates and potential job candidates perceive the decision-making process including during all communication stages (second party effects), evaluate how neutral outside observers perceive how second parties are treated (one type of third-party understanding), and determine likely external legal assessment such as via compliance with the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission et al., 1978) and Title VII of the Civil Rights Act of 1964, the legal standards for such decisions in the United States (a second type of third-party understanding). Importantly, these legal dimensions are considered somewhat differently between those with psychological backgrounds (e.g., Hough et al., 2001) and machine learning backgrounds (e.g., Barocas & Selbst, 2016). Both perspectives should be considered explicitly.

Meta-Components

The last category of components is labeled *meta-components*, because the following issues should be explored during the examination of all previous components. These are overarching concerns that apply before, during, and after the creation of machine learning models. They are also generally the most complex to evaluate.

The tenth component, cultural context, refers to evaluation of assumptions made during any of the prior steps relating to culture, which includes the cultures of developers, first parties, second parties, and third parties. Although a comprehensive list of cultural concerns is far outside the scope of this paragraph, we highlight four major cultural dimensions that every auditor should explore. First, legal defensibility is always culturally embedded such that local laws and regulations drive what a particular culture views as "fair." For our example in this article, we have focused on compliance with employment law in the United States, which itself focuses on validity and avoiding bias. However, even this focus is culturally embedded; in the European Union, for example, employment law focused more heavily upon applicant privacy rights (Bhave et al., 2020). Second, power differentials vary across cultures such that second and third parties in high power distance cultures may be less willing to express dissatisfaction, even privately, with decisions made by model developers (Smith et al., 1998). Yet an inability to measure dissatisfaction does not imply satisfaction; care must be taken in such cultures to draw accurate conclusions. Third, the importance of community voice, which in this context refers to how much second parties expect to be involved in

decision-making, varies widely by culture. In our focal example, workers in many European Union countries expect their work councils to represent their interests in employment decisions (Pudelko, 2006); they may even expect to approve the use of models for hiring. Fourth, auditors must consider the role and influence of the political systems in which model use and development is embedded. Even the best model will be ineffective if an authoritarian government can influence how the model is selectively used or ignored after implementation.

The eleventh component, respect, refers to adherence to an indicated guiding framework for ethics in all modeling and deployment decisions. As described earlier, there are many such ethical codes, like that of the APA (2017). It is the responsibility of an auditor to ensure that all decisions made in all other components adhere to a meaningful ethical code or codes clearly articulated by the model developer. Furthermore, auditors should consider if these ethical codes are sufficient to address all aspects of the model being used. In our focal example, there are already well-established professional frameworks for the evaluation of employment decisions, including those produced by machine learning models. As such, the auditor should ensure that the issues raised in such frameworks were considered at all stages.

Finally, the twelfth component, research designs, refers to the use of high-quality, informative, appropriate research designs in support of not only overarching claims but also all subclaims. This encourages auditors to consider sampling, experimental design choices, variables, analytic approaches, and interpretation in support of all data collection used for any purpose. In our focal example, we might imagine that the developers realized that they had never evaluated how job candidates felt about the algorithm they developed, so they conducted a small 10-person focus group to get reactions. This focus group should then be evaluated completely, as if it were a stand-alone research study, to ensure that the claims resulting from it are valid and appropriate. Every distinct data collection effort used to make a claim about an algorithm must be evaluated completely, in all relevant dimensions.

Recommendations for the Future of AI Auditing

Multidisciplinarity

Given the wide range of technical, ethical, legal, and practical issues that must be considered in the design and audit of AI systems, multidisciplinarity is required. Although we have focused more heavily on the psychological dimensions of audits, audits should integrate a psychological perspective on measurement and human characteristics, a machine learning and AI perspective on each major aspect of system design and function, and domain-specific knowledge relevant to the claims being made. For example, an audit of a predictive policing system based upon models using image data as features would be incomplete without the perspectives of experts in the psychology of behavior, experts in computer vision, experts in predictive modeling, and experts in policing. An audit of a system using drone-provided video data from disaster sites to inform triage practices would be incomplete without experts in triage and medical diagnosis. Including and exploring the assumptions and emphases brought by each perspective enhances the credibility of the audit overall and should not be dismissed.

Certification and Professionalization

The threat posed by siloed auditing is particularly salient as the number of professionals advertising AI auditing services continues to grow nearly as quickly as the use of AI itself. We anticipate a near future in which these auditors will themselves require auditing, in the form of certification; audit-the-auditor services will no doubt become popular. Existing regulatory frameworks that call for auditing, for example, Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (2021), provide few details on what an audit entails, and when details are provided, focus on issues like traceability and transparency, which fall on the machine learning side of audits as we have described them here. Narrow regulations like this incentivize similarly narrow and incomplete auditing procedures that can undermine public trust.

Collaboration

To build a world with high-quality algorithms deserving of the public's trust, a common language and spirit of collaboration is needed between psychologists, AI developers, and subject matter experts in the domains of capabilities being tested, operating as equal partners throughout the design, implementation, and monitoring of algorithmic systems. Just as data scientists have begun to bring psychological measurement into data science (Jacobs & Wallach, 2021), we must reciprocate to remain relevant. Psychologists and subject matter experts cannot wait until algorithmic systems are in use before contributing their expertise, evaluating algorithmic successes and failures long after implementation, just as algorithm developers cannot blindly build prediction engines without consideration of the contexts in which their models will be used. We need each other, at every step, to develop the best AIs possible; at this point, prediction systems built within disciplinary silos are incomplete at best and destructive at worst.

Conclusion

As the number of high-stakes algorithmic decisions continues to increase in modern society, we will only be able to fairly select the best employees, correctly identify violent criminals, save lives and beyond, by working together across the disciplinary barriers that our intellectual forebears created

yet we often unthinkingly reinforce and extend despite a changing world. We intend the present work both to serve as a useful organizing framework for psychologists and to represent an olive branch to technical and subject matter embedded disciplines. We encourage the use of such interdisciplinary audits as routine measures, used to support the design and development of algorithmic systems at all stages instead of merely being defensively deployed to avoid litigation and regulatory violations. Auditing is a process, not a product. We entreat future researchers across backgrounds to engage fully with it, to better equip society to meet the grand challenge of creating and managing valid, ethical, and useful AI.

References

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, *28*(4), 399–416. https://doi.org/10.1111/ijsa.12306

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

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code/ethics-code-2017.pdf

Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. *California Law Review*, *104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Berk, R. A. (2021). Artificial intelligence, predictive policing, and risk assessment for law enforcement. *Annual Review of Criminology*, *4*(1), 209–237. https://doi.org/10.1146/annurev-criminol-051520-012342

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Bhave, D. P., Teo, L. H., & Dalal, R. S. (2020). Privacy at work: A review and a research agenda for a contested terrain. *Journal of Management*, *46*(1), 127–164. https://doi.org/10.1177/0149206319878254

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478–494. https://doi.org/10.1037/0021-9010.74.3.478

Binns, R., & Gallo, V. (2020, July 20). *An overview of the Auditing Framework for Artificial Intelligence and its core components*. Information Commissioner's Office. https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

Boudana, S. (2016). Impartiality is not fair: Toward an alternative approach to the evaluation of content bias in news stories. *Journalism*, *17*(5), 600–618. https://doi.org/10.1177/1464884915571295

Brenner, F. S., Ortner, T. M., & Fay, D. (2016). Asynchronous video interviewing as a new technology in personnel selection: The applicant's point of view. *Frontiers in Psychology*, *7*, Article 863. https://doi.org/10.3389/fpsyg.2016.00863

Brockner, J., Fishman, A. Y., Reb, J., Goldman, B., Spiegel, S., & Garden, C. (2007). Procedural fairness, outcome favorability, and judgments of an authority's responsibility. *Journal of Applied Psychology*, *92*(6), 1657–1671. https://doi.org/10.1037/0021-9010.92.6.1657

Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification* [Conference session]. Conference on Fairness, Accountability and Transparency, New York, United States. http://proceedings.mlr.press/v81/buolamwini18a.html

Electronic Privacy Information Center. (n.d.). *AI and human rights*. Retrieved August 1, 2021, from https://epic.org/ai/

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures*. http://uniformguidelines.com/uniguideprint.html

Ford, D. K., Truxillo, D. M., & Bauer, T. N. (2009). Rejected but still there: Shifting the focus in applicant reactions to the promotional context. *International Journal of Selection and Assessment*, *17*(4), 402–416. https://doi.org/10.1111/j.1468-2389.2009.00482.x

Gilligan, C. (1995). Moral orientation and moral development. In V. Held (Ed.), *Justice and care* (pp. 31–46). Routledge. https://doi.org/10.4324/9780429499463-4 (Original work published 1987)

Gonzalez, M., Capman, J., Oswald, F., Theys, E., & Tomczak, D. (2019). "Where's the I-O?" artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, *5*(3). https://doi.org/10.25035/pad.2019.03.005

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports*, *21*(11), Article 116. https://doi.org/10.1007/s11920-019-1094-0

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management*, *16*(2), 399–432. https://doi.org/10.1177/014920639001600208

Holtz, B. C., & Harold, C. M. (2009). Fair today, fair tomorrow? A longitudinal investigation of overall justice perceptions. *Journal of Applied Psychology*, *94*(5), 1185–1199. https://doi.org/10.1037/a0015900

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1–2), 152–194. https://doi.org/10.1111/1468-2389.00171

Hutchinson, B., & Mitchell, M. (2019). *50 years of test (un)fairness: Lessons for machine learning* [Conference session]. Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, United States. https://doi.org/10.1145/3287560.3287600

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). *Towards accountability for machine learning datasets: Practices from software engineering and infrastructure* [Conference session]. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, United States. https://doi.org/10.1145/3442188.3445918

Jacobs, A. Z., & Wallach, H. (2021). *Measurement and fairness* [Conference session]. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, United States. https://doi.org/10.1145/3442188.3445901

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology*, *86*(5), 984–996. https://doi.org/10.1037/0021-9010.86.5.984

Kim, Y., & Ployhart, R. E. (2018). The strategic value of selection practices: Antecedents and consequences of firm-level selection practice usage. *Academy of Management Journal*, *61*(1), 46–66. https://doi.org/10.5465/amj.2015.0811

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.),

Proceedings of the 8th innovations in theoretical computer science conference. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Kobielus, J. (2021). How we'll conduct algorithmic audits in the new economy. *InformationWEEK*. https://www.informationweek.com/big-data/ai-machine-learning/how-well-conduct-algorithmic-audits-in-the-new-economy/a/d-id/1340299

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review of the effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, Article 106878. https://doi.org/10.1016/j.chb.2021.106878

Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. (2021). *Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts* (COM/2021/206, 52021PC0206). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). *Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation* [Conference session]. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York, United States. https://doi.org/10.1145/3351095.3372867

McDaniel, A., DiPrete, T. A., Buchmann, C., & Shwed, U. (2011). The black gender gap in educational attainment: Historical trends and racial comparisons. *Demography*, *48*(3), 889–914. https://doi.org/10.1007/s13524-011-0037-0

McLaughlin, M., & Castro, D. (2020). The critics were wrong: NIST data shows the best facial recognition algorithms are neither racist nor sexist. *Information Technology and Innovation Foundation*. https://itif.org/publications/2020/01/27/critics-were-wrong-nist-data-shows-best-facial-recognition-algorithms

McPhail, S. M. (2005). Auditing selection processes: Application of a risk assessment model. *The Psychologist Manager Journal*, *8*(2), 205–221. https://doi.org/10.1207/s15503461tpmj0802_10

Meade, A. W., Behrend, T. S., & Lance, C. E. (2009). Dr. StrangeLOVE, or: How I learned to stop worrying and love omitted variables. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 89–106). Routledge/Taylor & Francis.

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(2), 192–205. https://doi.org/10.1111/j.1754-9434.2010.01223.x

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Miller, K. (2021). *Renata Avila: Reclaiming AI's superpowers for collective good*. Stanford HAI. https://hai.stanford.edu/news/renata-avila-reclaiming-ais-superpowers-collective-good

Moskop, J. C., & Iserson, K. V. (2007). Triage in medicine, part II: Underlying values and principles. *Annals of Emergency Medicine*, *49*(3), 282–287. https://doi.org/10.1016/j.annemergmed.2006.07.012

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–36. https://doi.org/10.1145/3359221

Norman, W., & MacDonald, C. (2004). Getting to the bottom of "triple bottom line." *Business Ethics Quarterly*, *14*(2), 243–262. https://doi.org/10.5840/beq200414211

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council of Artificial Intelligence* (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Pudelko, M. (2006). A comparison of HRM systems in the USA, Japan and Germany in their socio-economic context. *Human Resource Management Journal*, *16*(2), 123–153. https://doi.org/10.1111/j.1748-8583.2006.00009.x

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing* [Conference session]. FAT* '20: Conference on fairness, accountability, and transparency, Barcelona, Spain. https://doi.org/10.1145/3351095.3372873

Roth, P. L., Bobko, P., Mabon, H., Anderson, N., Ones, D. S., & Viswesvaran, C. (2002). Utility analysis: A review and analysis at the turn of the century. In N. Anderson & D. S. Ones (Eds.), *Handbook of industrial, work & organizational psychology: Volume 1* (pp. 383–384). Sage Publications.

Sackett, P. R., Corte, W. D., & Lievens, F. (2008). Pareto-optimal predictor composite formation: A complementary approach to alleviating the selection quality/adverse impact dilemma. *International Journal of Selection and Assessment*, *16*(3), 206–209. https://doi.org/10.1111/j.1468-2389.2008.00426.x

Sale of Automated Employment Decision Tools. (2020). *Sale of automated employment decision tools, no. 1894*. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID

Schroepfer, M. (2021, March 11). *Teaching fairness to machines*. Facebook Technology. https://tech.fb.com/teaching-fairness-to-machines/

Smith, P. B., Dugan, S., Peterson, A. F., & Leung, W. (1998). Individualism: Collectivism and the handling of disagreement. A 23 country study. *International Journal of Intercultural Relations*, *22*(3), 351–367. https://doi.org/10.1016/S0147-1767(98)00012-1

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures*. https://www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf

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*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

The Public Voice. (2018). *AI universal guidelines explanatory memorandum*. https://thepublicvoice.org/ai-universal-guidelines/memo/

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021, January 28). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. *Personnel Assessment and Decisions*, *7*(2). https://doi.org/10.31234/osf.io/6gczw

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*(1), 4–70. https://doi.org/10.1177/109442810031002

Vanian, J. (2021). Federal watchdog says A.I. vendors need more scrutiny. *Fortune*. https://fortune.com/2021/07/13/federal-watchdog-a-i-vendors-need-more-scrutiny/

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Received August 5, 2021
Revision received November 16, 2021
Accepted November 20, 2021 ■