Published in final edited form as:
Perspect Psychol Sci. 2023 September ; 18(5): 1062–1096. doi:10.1177/17456916221134490.

A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health

Adela C. Timmons[1,2], **Jacqueline B. Duong**[1], **Natalia Simo Fiallo**[3], **Theodore Lee**[3], **Huong Phuc Quynh Vo**[4], **Matthew W. Ahle**[2], **Jonathan S. Comer**[3], **LaPrincess C. Brewer**[5,6], **Stacy L. Frazier**[3], **Theodora Chaspari**[4]

[1] University of Texas at Austin Institute for Mental Health Research

[2] Colliga Apps Corporation

[3] Florida International University

[4] Texas A&M University School of Computer Science and Engineering

[5] Department of Cardiovascular Medicine, May Clinic College of Medicine, Rochester, Minnesota, United States

[6] Center for Health Equity and Community Engagement Research, Mayo Clinic, Rochester, Minnesota, United States

Abstract

Advances in computer science and data analytic methods are driving a new era in mental health research and application. Artificial intelligence (AI) technologies hold the potential to enhance the assessment, diagnosis, and treatment of people experiencing mental health problems and to increase the reach and impact of mental health care. However, AI applications will not mitigate mental health disparities if they are built from historical data that reflect underlying social biases and inequities. AI models biased against sensitive classes could reinforce and even perpetuate existing inequities if these models create legacies that differentially impact who and how effectively a person is diagnosed and treated. The current article reviews the health equity implications of applying AI to mental health problems, outlines state-of-the-art methods for assessing and mitigating algorithmic bias, and presents a call to action to guide the development of "fair-aware AI" in psychological science.

Keywords

Artificial intelligence; fair-aware; bias; mental health equity

Since the beginning of clinical psychological science in the 1890s, the field has made important advancements, publishing thousands of research articles and developing over 500 evidence-based mental health therapies for more than 150 different classified mental health disorders (American Psychiatric Association, 2013; APA Presidential Task Force

Correspondence regarding this article should be addressed to Adela C. Timmons, Institute of Mental Health Research, Department of Psychology, University of Texas at Austin, 108 E Dean Keeton St, Austin, TX 78712, adela.timmons@austin.utexas.edu.

on Evidence-Based Practice, 2006; Benjamin Jr., 2005). As stated in the 2012 *American Psychological Association Resolution on the Effectiveness of Psychotherapy*, mental health treatments are generally effective across a range of mental health problems (e.g., American Psychological Association, 2012; Cuijpers & Gentili, 2017; Hofmann et al., 2012; Weisz et al., 1995). However, despite considerable progress in clinical research and treatment, critical targets for improving the accessibility and effectiveness of mental health treatments remain. Even with the recent proliferation of evidence-based mental health therapies, mental health problems continue to affect 20% of the population and cost the global economy $2.5 trillion annually in health care, lost work productivity, disability, and mortality (Lancet Global Health, 2020; National Institute of Mental Health, 2022). The COVID-19 pandemic has exacerbated this already heavy toll, resulting in a widely recognized worldwide mental health crisis (e.g., Lancet Global Health, 2020; Lancet Global Health Covid 19 Mental Disorders Collaborators, 2021; Radfar et al., 2021).

Causes of shortfalls in mental health care, although complex, can be understood as generating from two broad categories: reach and impact. Reach includes the ability to access and engage with therapy. Fewer than half of people in the US who need mental health services receive them (National Institute of Mental Health, 2022). Factors such as lack of health insurance and lack of financial resources; lack of providers, lack of providers identifying with minoritized and marginalized groups, regional workforce disparities, and provider burnout; time and logistical constraints related to childcare, work, and transportation; prioritization of more urgent needs, such as food, safety, or housing; cultural differences in conceptualizing mental health problems and lack of alignment between available interventions and cultural values; and social stigma limit an individual’s capacity to obtain treatment (Alegría et al., 2018; Carbonell et al., 2020). Again, the COVID-19 pandemic has intensified these problems, resulting in long waitlists for remote mental health services and decreased opportunities to access in-person care (Ornell et al., 2021; World Health Organization, 2020). Impact, in contrast, refers to the effectiveness of the treatments that are administered. A large body of work demonstrates the overall efficacy of psychotherapy, but for some people, treatment works only partially or not at all. Approximately 25 to 50% of those people who receive psychotherapy show no benefit, and an estimated 50 to 70% of people retain their clinical diagnoses at the end of treatment (American Psychological Association, 2020; Hofmann et al., 2012; Rozental et al., 2019).

Limitations in the reach and impact of mental health care have even greater implications for those people from resource-limited backgrounds, racial and ethnic minoritized groups, and those with severe mental health problems (e.g., Maura & de Mamani, 2017; McGuire & Miranda, 2008; Miranda et al., 2008). Psychological research has historically been designed by and tested on primarily White and high-income people, limiting the socio-cultural sensitivity, inclusivity, and responsivity of our treatments and skewing our conceptualizations of the etiology and treatment of mental health problems toward the perspectives of the majority culture (Roberts et al., 2020). Compared to their non-Hispanic White counterparts, members from minoritized racial and ethnic groups are less likely to seek and obtain mental health treatment and are more likely to receive poor quality services when treated (Abe-Kim et al., 2007; American Psychiatric Association, 2017; Miranda et al., 2008; Sussman et al., 1987; Zhang et al., 1998). The adverse effects of the

COVID-19 pandemic have further exacerbated these social, racial, and ethnic mental health inequities, with those from resource-limited and racial and ethnic minoritized backgrounds experiencing greater exposure to the SARS-CoV-2 virus, higher mortality rates, fewer opportunities to work remotely, greater loss of income, and higher levels of COVID-19-related stigma and discrimination (e.g., Hooper et al., 2020; McKnight-Eily et al., 2021; Miconi et al., 2021; Perry et al., 2021; Purtle, 2020; Saltzman et al., 2021; Sneed et al., 2020). Differential exposure to the stress and trauma related to COVID-19, combined with barriers to accessing mental health services for historically underserved and under-resourced populations, highlight the need to develop effective and socio-culturally sensitive treatments to support diverse people. To reach and provide mental health services and supports to all who need them, the field has become increasingly focused on developing better ways to deliver mental health treatments; many such developments now focus on technological innovation as a promising tool to address mental health disparities (Comer & Barlow, 2014; Kazdin & Blase, 2011). These technologies use the internet, smartphones, wearables, pervasive computing, and artificial intelligence (AI) to deliver and enhance mental health interventions.

The Promise of AI: A New Era for Psychological Science

AI is defined as the theory and development of computer systems that can perform human cognitive functions, such as visual perception, speech recognition, decision-making, and language (Joiner, 2018). When applied to the science of mental health, AI pertains to using machines to assist humans in diagnosing and treating mental health problems. The application of AI to mental health problems can take various forms (Dwyer et al., 2018; Graham et al., 2019; Shatte et al., 2019). Recent applications of AI include using biological markers to assist in making psychological diagnoses (e.g., Battista et al., 2020; Cai et al., 2018; Erguzel et al., 2016; Patel et al., 2015) or using medical records to predict which people are most vulnerable to relapse, will need readmission, or require more extensive treatment (e.g., Arun et al., 2018; Choi et al., 2018; Dobias et al., 2022; Fernandes et al., 2018). Other work focuses on using AI, smartphones, and wearable devices to monitor functioning remotely (e.g., sensing stress from biological signals; Faedda et al., 2016; Jacobson & Bhattacharya, 2022; Wahle et al., 2016; Yamamoto et al., 2018) and to send interventions via smart devices (e.g., a text prompting a person to complete a meditation) in naturalistic settings and specific moments of need (e.g., Carpenter et al., 2020; Juarascio et al., 2018; Nahum-Shani et al., 2018; Perski et al., 2021).

AI and related mental health technologies putatively enhance traditional services and delivery systems by helping create effective, affordable, and scalable automated and assistive technologies (Davenport & Kalakota, 2019; Noah et al., 2018; Queirós et al., 2018; Warren et al., 2018). Furthermore, AI and digital health interventions hold great potential as tools to address disparities in mental health treatment reach and impact (Bravo et al., 2014; National Academies of Sciences, 2016; Varshney, 2007). According to a 2021 survey of adults by the Pew Research Center, roughly 76% of people from resource-limited backgrounds own smartphones (Vogels, 2021). People from resource-limited backgrounds are among the fastest rising adopters of smartphones (Anderson-Lewis et al., 2018) and are more likely than their high-income peers to rely on a phone, rather than a computer or another device, to access the internet for work, school, or entertainment purposes (Vogels, 2021). Smart devices are a cost-effective way to deliver mental health services and support (Iribarren et al., 2017; Price et al., 2014) and from a scientific perspective, can be used to collect ecologically valid time series data for testing key questions related to clinical intervention and therapeutic change. For example, AI modeling and related big data methodologies can be harnessed to (1) identify mechanisms underlying therapy response and their common factors to create unified mental health protocols (Barlow et al., 2017); (2) identify group versus individual therapy response predictors to adapt treatments to each individual through a model of precision medicine (Johnson et al., 2021); and (3) refine our methods of treatment administration to optimize precisely when, where, and how interventions are delivered (Hardeman et al., 2019). Although AI-based mental health technologies are not intended to replace traditional therapies and are not appropriate for all situations, such technologies could lead to a more flexible model capable of delivering effective mental health support to a greater percentage of the population.

Dangers in the Road Ahead: Will AI Result in the Perpetuation of Past Inequities?

As enthusiasm about the potential of AI to improve the reach and impact of mental health care grows, new technologically-based mental health treatments are being developed at breakneck speed (Bohr & Memarzadeh, 2020; Hindley et al., 2022; National Institute of Mental Health, 2017; Shaheen, 2021; Torous et al., 2020). Much of this work is motivated by the potential of AI to mitigate existing disparities in mental health. Yet, accomplishing this vision requires critical foundational research and development, careful consideration of the ethical issues involved, and a thorough understanding of how mental health-related AI technologies impact the people using them. What happens if an algorithm detects or predicts a risk event requiring a clinician to break confidentiality and make a report? Or, what if an algorithm gets a diagnosis or prediction wrong? Applying AI to clinical psychological problems requires researchers and developers to carefully consider issues such as privacy and confidentiality; harm prevention; risk and emergency monitoring and response; the development and adoption of regulatory frameworks for AI use; the identification and prevention of the misapplication of algorithms developed; an understanding of the implications of replacing established services; respect and protection of autonomy; transparency in building and applying algorithms; and understanding the long-term implications of algorithmic use and misuse (e.g., Fiske et al., 2019; Fulmer et al., 2021; Noor, 2020).

Of particular and often overlooked importance is how algorithm development can lead to biased applications that reinforce and perpetuate past societal inequities (Brewer et al., 2020). The proliferation of available data and technological methods for extracting insights from them might give the impression that such technologies are unbiased. However, just as humans' past experiences and personal values influence their decision-making in biased ways, algorithms built *by people* on data collected *by people* are also subject to bias. Furthermore, the values and goals of algorithm developers may or may not coincide with users' values and may sometimes run contradictory to the public good. For example, AI applications commonly employed on news and social media websites utilize users' past activity to determine what content is presented to them in the future. These models are often optimized to increase clicks, a metric that typically translates to financial gains for the businesses deploying the algorithms. Users may assume that the algorithm is explicitly built to learn their interests and may not fully understand the repercussions of interacting with the system. Although curated content may be a secondary benefit, the algorithm's primary goal is to increase clicks. Thus, such an algorithm may predict what a person will click on but may fail to capture that person's other values, such as reading diverse and accurate content. Utilizing AI models based on flawed assumptions or incongruent values (i.e., a mismatch between developers' and users' interests) can lead to unintended impacts that can compound over time. Specifically, biased models can create vortexes of distorted reality that proliferate exponentially and infiltrate more aspects of daily life (e.g., as seen with the recent impact of inaccurate political content and false information on society; Ali & Hassoun, 2019; Kreps et al., 2020; Mohseni & Ragan, 2018).

Biased algorithms and their problematic outcomes threaten the ethical use of AI-built models for mental health. Regrettably, our field has a long and well-documented history of biased research that has both caused and contributed to disparities in mental health. This legacy is evidenced by the systematic underrepresentation of diverse people in research samples and the underrepresentation of research scientists from diverse backgrounds, especially scientists in leadership positions (Hall, 2006; Henrich et al., 2010; Huff, 2021; Jones, 2010; Turpin & Coleman, 2010). As we develop AI applications for mental health, we must take responsibility for preventing their misapplication. Indeed, AI applications hold great potential to deliver improved mental health services and support to a greater percentage of the population—especially underserved and disadvantaged populations. Yet, if we simply capture more data and use more advanced analytic strategies to analyze that data without examining and eliminating underlying biases, then we will perpetuate existing historical biases and their resulting inequities. Although the problem of bias pertains to all psychological research and science using inferential statistics, AI applications to mental health warrant special consideration. These applications provide a mechanism to scale mental health services at an unprecedented level, which presents great opportunity and risk. It is well-recognized that AI is one of the next great waves of innovation in our field (e.g., Dwyer et al., 2018), with the first applications now being implemented and increasing applications to be employed in the coming years. Thus, our field has an opportunity to set a new trajectory and prevent the repetition, perpetuation, and exacerbation of past mistakes.

This chance to avoid the misapplication of AI converges with a broader social movement, reckoning, and awakening relating to diversity, equity, inclusion, and justice in the mental health field and, more broadly, the nation (Roberts & Rizzo, 2021; Williams, 2019). This movement toward diversity, equity, inclusion, and justice in mental health is evidenced by the recent publication, *Call to Action for an Antiracist Clinical Science* (Galán et al., 2021) and other articles that call for the dismantling of oppression in psychological science (Andoh, 2021; Buchanan & Wiklund, 2020; Byrd et al., 2021; Hochman & Suyemoto, 2020). In the current article, we present a broad integrative review to serve as a general framework, to increase awareness in the field, and direct psychological scientists to the work currently being conducted by engineers and the burgeoning empirical science of equitable AI so that psychologists incorporate these developing standards within their own future AI applications. Given the very broad nature of the topic and the variety of different applications and scenarios it encompasses, it is not possible to include definitive answers to all questions relating to fairness and justice within the AI field or to provide detailed guidance on every bias evaluation and mitigation strategy. Rather, we aim to point psychological scientists to the broader fair-aware AI field so that evaluation and mitigation techniques are consistently employed moving forward, as is appropriate for each researcher's specific circumstance.

What Is Bias and How Can It Influence AI?

Bias, an innate and unavoidable aspect of life, is the tendency to favor or exhibit prejudice against something or someone over others. Biases can take many forms, including social bias (i.e., prejudice against a person or group), cognitive bias (i.e., systematic errors in thinking and information processing), and statistical bias (i.e., the tendency to overestimate or underestimate the parameters of a population (Sica, 2006). All people have biases, which are neither inherently good nor bad, as biases reflect the way the brain unavoidably learns, functions, and processes information (Sali et al., 2018). Biases help people make sense of the world around them and can be adaptive and beneficial, such as when one chooses to eat healthy foods (e.g., salad over pizza) or avoid dangerous experiences (e.g., walking next to a cliff). From a biological standpoint, biases in cognition and behavior serve as evolutionary heuristics that help improve decision-making (Johnson et al., 2013). However, biases are based on inaccurate assumptions and stereotypes, rather than accurate information about a group or actual knowledge of an individual, can lead to misguided decision-making and systematic discriminatory practices (De Houwer, 2019; Delgado-Rodriguez & Llorca, 2004).

Although some biases are evolutionarily hardwired (e.g., error management theory; Haselton & Buss, 2000), many biases are learned through early socialization and result from upbringing (Hitlin & Pinkston, 2013; Sali et al., 2018). Our lived experiences make up schemas and categories of knowledge that help us interpret and understand the world (Piaget, 1952). These organizational mental structures, influenced by individuals' experiences, guide their decision-making, behaviors, and interactions. In areas where we have little or no experience, we develop "gaps" in our cognitive schemas (Hefner, 2002). This limited representation of other people's life experiences in our cognitive schemas may lead us to make systematically prejudiced inferences, decisions, and judgments. Additionally, our narrow schemas can lead to functional fixedness (German & Defeyter, 2000), which prevents us from seeing and using information in new and nontraditional ways. This cognitive rigidity and overuse of traditional, expectancy-based thinking threatens the accuracy and utility of our scientific endeavors (Harley, 2017).

Biases can take explicit and implicit forms; explicit biases are intentional and controllable beliefs or attitudes toward groups, whereas implicit biases refer to our unconscious associations and preferences represented in our cognitive structures (Yarber, 2022). Implicit association tests (IATs) demonstrate the existence of these "gaps" through timed computer tasks that reveal peoples' unconscious tendencies to identify particular social identities (e.g., gender, ethnicity, race, age, sexuality) more quickly and automatically as either good or bad or as linked to specific qualities (e.g., female with caretaking; men with science; Greenwald & Krieger, 2006). Although both explicit and implicit biases are problematic, implicit biases are particularly insidious in their impact on society. Implicit biases become harmful when they limit an individual's or group's experiences and opportunities or cause an individual to be disadvantaged or experience discrimination due to their social identity. Implicit bias can lead to unfair decisions in the workplace (e.g., who is selected for employment or promotion), school (e.g., which students are penalized for misbehavior, placed in advanced classes, or admitted to prestigious schools), and health care (e.g., which treatment options we make available to which patients; Jost et al., 2009). Unfortunately, implicit biases are often inaccurate and result in systematic discrimination against individuals and groups (Payne & Hannay, 2021; Payne et al., 2017). Moreover, these biases can have significant downstream effects regarding the scientific questions we ask, the initiatives we prioritize, the laws we enact, and the social structures we form.

Just as underlying cognitive biases impact how build and interact with society at large, contributing to systemic inequities, statistical biases underlying the development of algorithms can and will result in similar inequities. In statistical procedures, bias occurs through unrepresentative sampling or inaccurate measurement (Delgado-Rodriguez & Llorca, 2004; Johnson et al., 2022). Like human biases that result from inadequate exposure to diverse life experiences, AI biases arise from insufficient exposure to "experience" in the form of representative and accurate data. As AI algorithms are continually developed and employed, they will create systems and networks paralleling that of schemas in the human brain and will be subject to similar limitations. Undoubtedly, the AI will reflect such biases if we continue to build from a foundation of historical and contemporary biases. Further, beyond reflecting existing biases in our society, such algorithms, if used to make important decisions regarding health, law, or access to opportunities or resources, may solidify, reinforce, perpetuate, and amplify inequities. When the AI decision-making process systematically biases decisions against one group (e.g., overestimation of recidivism), it impacts that group's outcomes (e.g., harsher prison sentences, decreased opportunities, decreased access to resources, and decreased mental health), which then impacts the future decision-making of the algorithm (i.e., confirmation and amplification of the biased decisions). Ultimately, biased algorithmic decisions reflect more than isolated computations and can contribute to building social structures that create legacies with long-lasting consequences (Cummings, 2021).

The Science of AI: A Broad Overview of AI Methodology

Traditionally, AI has not been included as a standard part of training in psychological science doctoral programs. But in recent years, the field has come to recognize the importance of data science, big data, AI, and machine learning in psychological research and application (Blease et al., 2021; Graham et al., 2019; Luxton, 2014). The proliferation of AI in psychological science is evidenced by the increasing number of targeted faculty hires, newly opened data science centers, courses and concentrations offered, and training programs that explicitly focus on AI and interdisciplinary data science models (e.g., Cooper, 2020; Irizarry, 2020; Kusters et al., 2020). Although a comprehensive introduction to the methods and procedures of AI are outside the scope of this paper (see Lantz (2019)or

Dwyer et al. (2018) for further reading), a brief description of AI methods is provided as a framework for grounding our discussion about how bias in AI impacts clinical science.

The primary goal of AI is to use computers rather than humans to make decisions or inferences. Relegating decision-making to a computer, if the algorithm works well and achieves its intended purpose, results in three advantages: (1) increased accuracy in making decisions, (2) the generation of new insights that human minds would be unable to independently find or recognize, or do so as quickly, efficiently, and inexpensively, and (3) enhanced ability to "scale," or to implement a decision-making task quickly and cost-effectively on a population level. To date, AI has been deployed most commonly in industry settings (e.g., Healthcare, retail and e-commerce, food tech, banking and financial services, logistics and transportation, travel, real estate, entertainment and gaming, manufacturing; Techjury.net, 2019). For example, a common application deployed in advertising involves using AI to contextually advertise and behaviorally target niche consumer populations with specific ads (IBM Watson Advertising, 2021). Such AIs can generate insights (e.g., a person clicked on an ad for a crib and thus should receive ads for related items, such as a stroller or baby bottles) and implement decisions (e.g., show an ad) at a scale (e.g., millions of consumers), speed (e.g., seconds), and cost (e.g., $1 per click) that human sales teams cannot match. Unlike traditional psychological research, which focuses on answering scientific questions, business-related AI aims to build cost-effective and efficient *tools* for achieving specific goals, such as reaching more people via advertising to generate sales and increase profits.

AI methods are generally classified into either unsupervised, supervised, or reinforcement learning techniques (Delua, 2021). Unsupervised learning refers to identifying the underlying associations or subgroups in the data that were not previously labeled (i.e., unknown outcomes). Supervised learning, in contrast, refers to algorithm development where data with known outcomes are used to generate rules or associations between an input and an output. After the decision-making algorithm has been created and validated using known outcomes, it is applied to new or "unseen" data for which outcomes are unknown. The final category, reinforcement learning, allows models to improve, even while being implemented, by incorporating user feedback as an outcome into the system and continually updating the model (Qi et al., 2021). In supervised and reinforcement learning models, rules are designed to optimize accuracy in making decisions about the known outcomes. In this paper, we focus on supervised methodologies. The development of supervised algorithms generally follows a standardized procedure that includes a training phase, where the algorithm is built, and a testing phase, where the algorithm is evaluated based on its predictive performance on unseen data. The first step in the training phase is identifying the task of interest. For instance, a researcher might wish to determine who is at risk for rehospitalization after admission to the psychiatric emergency room (i.e., the outcome of interest). Once the task is identified, the developer collects data upon which to develop and validate the algorithm. Data to be collected must include "features," also known as predictor or input variables, and "labels," (also referred to as ground truth or outcome variables; Krishnapuram et al., 2005; Topol, 2020). The features are ideally chosen to reliably detect or predict the outcome of interest. Labels are considered "ground truth" because they define the state of truth that directs the optimization of the algorithm. As discussed in the following sections, decisions about operationalizing algorithm labels have important implications for algorithmic bias. In our example above, an algorithm developer might collect data from hospital records (e.g., past assessments, diagnoses, and number of prior hospitalizations) to use as features and then collect data on who is re-hospitalized 6 months later to use as labels.

Once the data are obtained, the developer selects a model to build the algorithm. The choice of model varies by the application, goal, sample size, type of data to be used, and features of the outcome to be detected or predicted. Different models are appropriate for different cases, and each model has advantages and disadvantages associated with its use. Examples of commonly used models include neural networks, decision trees, and support vector machines (Hearst et al., 1998; Quinlan, 1990; Warner & Misra, 1996). Methods for building algorithms vary across models but are generally built by iterating through feature combinations (e.g., if the number of prior hospitalizations is greater than 2 and the depression test score is greater than 15; if the number of prior hospitalizations is greater than 3 and the assessment depression score is greater than 20) or by transforming features so that their combination yields a reliable estimate of the outcome of interest. After examining a variety of combinations, the algorithm selects an equation that maximizes performance on the known outcome. Different models employ different criteria for making selections, such as choosing feature combinations that maximize accuracy, R-squared, kappa, F1 score, sensitivity, specificity, area under the curve, or other related metrics. A central benefit of AI is that a computer can assess many feature combination equations beyond what the human mind could accurately and efficiently achieve. Although some models yield easily understandable results, such feature combinations can sometimes be highly complex. Given the rapid advancement of “black-box” methodologies (i.e., a system whose inner workings are unknown or difficult to interpret), recent efforts have also concentrated on building interpretable algorithms that promote transparency and trustworthiness by explaining how the algorithmic decisions are made (e.g., Dave et al., 2020; Khedkar et al., 2020; Plous & Holm, 2020).

After a model is developed in the training phase, the researcher next evaluates the model’s performance in the testing phase. Algorithms built on specific data will “overfit” the model to the peculiarities of that data. Thus, all models built on training data will overestimate model performance and, in turn, must be validated on unseen testing data before implementation. In AI-based research, appropriate validation techniques must be employed to prevent type I errors, spurious conclusions about associations in the data, and the overestimation of model performance. Validation using unseen data can be conducted in multiple ways; a standard procedure, especially in the initial model-building stage, is to withhold a certain percentage of randomly sampled data (most typically around 10–30%) during model training and then evaluate performance on the withheld data. The data used to build the model are called the *training data,* and the data used to evaluate the model are called the *test data*. The model built during the training phase is then applied to the test data and used to generate predictions. Model performance is subsequently assessed by comparing the developer-provided labels to the decisions generated by the algorithm and calculating various indices of performance (e.g., kappa, sensitivity, specificity, area under the curve, F1 score). At a minimum, algorithms should perform better than chance accuracy (e.g., 0.5 sensitivity for a binary classification task with balanced classes), and for most applications, should perform significantly better than chance accuracy (e.g., 0.7–0.8 sensitivity for a binary classification task with balanced classes) to be useful, although the standards for algorithmic accuracy will differ depending upon the application being tested (e.g., tolerance of false negatives may differ depending on the severity of the diagnosis being predicted). Of note is the consideration of class sizes; the accuracy estimates for sparse events will be high if the algorithm always guesses the absence of the event, so metrics such as kappa that account for base cell sizes are typically necessary to ensure that the algorithm performs above chance levels.

Beyond dividing the data into one training dataset and one testing dataset, a more sophisticated cross-validation approach, termed *k-fold cross-validation*, is often employed. This method uses the same principles as above, but instead of using two samples, the model selects a new training set and test set for *k* number of folds. Specifically, the model randomly selects a training and test set (fold 1) and computes the result, randomly selects a new training and new test set (fold 2) and computes the result, and continues across all "*k*" folds (Rodriguez et al., 2010). The results are then combined at the end of the procedure. Although any number of folds can be used, a common choice is 10 folds; fewer folds may be preferable for small sample sizes. Another validation method, commonly used in mental health applications, is the *leave-one-participant-out cross-validation* method, where data from one person are withheld per fold. The model is then tested on the person left out, and the process repeats for each person in the dataset. The choice between *k-fold* and *leave-one-participant-out cross-validation* depends on what estimate of accuracy is of interest. *K-fold cross-validation* is typically used to predict information when we have some (randomly chosen) labeled data from a participant. *Leave-one-participant-out* is typically used to predict information for a new person. In datasets where repeated observations are collected (e.g., daily mood reports over a month), *leave-one-participant-out cross-validation* also ensures that performance is not inflated by sampling non-independent observations. Because each dataset contains peculiarities specific to that sampling (e.g., time of year, location of sampling), models built via cross-validation will also overfit the data, albeit less so than if cross-validation was not employed. Thus, after validating performance via cross-validation, the model should be tested on completely unseen data, i.e., an entirely new dataset, to verify performance before wide-scale implementation and dissemination.

Although the above summary provides a general outline of AI procedures, AI researchers are continually building new methodologies for improving model accuracy and expanding the bounds of what is possible in AI. For example, the Caret package in R (Kuhn, 2009) allows researchers to use automated selection processes to choose the models, features, and parameter values employed to maximize algorithm performance. Other methodologies, such as boosting, bagging, and stacking (Sutton, 2005), collectively called "ensemble" techniques, combine multiple models and learning methods to enhance performance. Recent work has also focused on building personalized models, which aim to build algorithms specific to each person using the system. Some applications combine personalized modeling with reinforcement learning to improve model accuracy for individuals over time (den Hengst et al., 2020). Models that are capable of improving without direct human instruction are specially referred to as *machine learning models* because they can "learn" with the accumulation of "experience" over time (Ngiam & Khor, 2019). Other work combines

AI and human-based methods. In this framework, the AI provides a structure for clinical decision-making, but the final decision is made by a human. This framework has been implemented in various mental and physical health applications, such as self-injury risk (Ammermann et al., 2020; Brandmaier et al., 2013; Djulbegovic et al., 2018). Such methods can increase accuracy and efficiency in making clinical decisions while also maintaining interpretability.

How Bias in AI Could Impact the Mental Health Field

Psychologists are increasingly recognizing the potential of AI to improve diagnostic accuracy, generate insights through the computational power of automated analysis, and reduce human burden, resulting in more scalable and accessible mental health care. Although AI has great potential to enhance the field of mental health, it also holds the potential to be misapplied, either via algorithms that are biased against specific groups or social identities; the applications of algorithms to groups for whom model performance has not been validated; or through the inadequate application of cross-validation techniques that could lead to inflated type I errors and thus erroneous conclusions regarding the associations observed in the dataset. Applications of AI in mental health can be summarized into five general categories, including diagnosis, intervention, engagement and adherence, maintenance and monitoring, and prediction. In the following sections, we discuss the potential uses of AI for each category. For a more extensive review of how AI can be applied in clinical science, see Dwyer et al. (2018) and Shatte et al. (2019).

Diagnosis and Clinical Classification.

Diagnosis refers to the classification of mental health symptom sets into categories used to direct treatment conceptualization and course. Diagnosis is the beginning of the person’s journey to treatment; thus, accurate diagnosis is imperative for successful treatment. Traditionally, diagnosis is made by humans and is usually based upon structured clinical interviews, observation, questionnaire-based assessments and testing, and medical record review. AI applied to clinical classification could be used to make diagnostic decisions without a human or to aid diagnosis by making classification faster or more accurate. AI may also be a valuable tool for understanding heterogeneity within psychiatric conditions by helping to discover subgroups within symptom sets, predictors of subgroup membership, and associations between subgroups and treatment outcomes (Feczko et al., 2019; Lombardo et al., 2019). Even in best practices when multiple sources of standardized and validated methods are used to make diagnoses, human bias often influences diagnostic decisions. Clinicians’ biases and social attitudes have been shown to affect who is diagnosed with what disorder, such as women being over-diagnosed with Borderline Personality Disorder or people from racial and ethnic minoritized groups being over-diagnosed with Conduct Disorder, even when the clinical presentations or case vignettes of the groups are identical (Fadus et al., 2020; Gara et al., 2019; Garb, 2021; Warner, 1979; Widiger & Spitzer, 1991). Like bias resulting from human-based diagnostic decisions, bias from AI-based diagnostic decisions has important implications for mental health equity. If certain populations are systematically underdiagnosed, they may become disadvantaged because they cannot access mental health services or receive insurance benefits (Gianfrancesco et al., 2018; Seyyed-

Kalantari et al., 2021). Conversely, systematic over-diagnosis of specific groups could lead to stigmatization, contribute to unfair and inaccurate perceptions of that group, or result in over-pathologizing normative or culturally grounded behaviors (Noor, 2020; Obermeyer et al., 2019).

Intervention.

After a diagnosis or classification is made, a course of treatment is typically initiated with the goal of improving or remitting symptoms related to a psychological problem. Treatment decisions include who will receive treatment, what treatment will be administered, and what level of support will be provided (e.g., outpatient treatment versus hospitalization, frequency of sessions, duration of treatment). AI-based mental health applications are typically designed to inform treatment decisions or make decisions in the stead of a human. Such AI methods aim to make therapy more effective or to automate treatment administration to make it less expensive or scalable to more people. Some algorithms focus on identifying which people should receive specific intervention modules or the ideal timing or order of the administered intervention modules. Other applications use smartphones and wearables to send interventions remotely and automatically ((JITAIs; Liao et al., 2018; Nahum-Shani et al., 2018; Wang & Miller, 2020) or use reinforcement learning to identify which techniques are working for a specific person, assign and send the most effective interventions more frequently, or develop individualized treatment programs. Importantly, AI therapies developed from homogenous samples may work well for the dominant group but may not work for those who are not adequately represented in the data. As new AI applications are developed, it will be critical to validate treatments in diverse populations to mitigate “digital divides” in mental health and to ensure that all people benefit from technological advances in treatment.

Engagement and adherence.

Effective intervention requires people to engage with therapy and adhere to the prescribed treatment protocol. Treatment engagement refers to people’s sustained participation in treatment (Dixon et al., 2016) whereas treatment adherence refers to whether the person’s behavior corresponds to clinician recommendations (Martin et al., 2005). Both adherence and engagement are linked to better treatment outcomes (Dixon et al., 2016; Donkin et al., 2011). Initial research suggests that technology may be an effective tool for promoting both engagement and adherence (Barello et al., 2016; Christie et al., 2021; Clough & Casey, 2011). With smartphones and computers, intervention content can be accessed and reviewed from home, homework assignments and completion reminders can be sent electronically, and people can complete symptom assessments quickly and easily. Moreover, reinforcement learning can be harnessed to determine when and under what conditions (e.g., when at home in the evening and looking at social media) people are more likely to complete homework assignments or respond to phone-based prompts. Another exciting technological application is the use of gamification, rewards, and point-scoring to encourage participation in treatment (Fleming et al., 2020; Pramana et al., 2018). Research to date indicates that people from racial and ethnic minoritized backgrounds, older adults, individuals with more severe symptoms, and people with low levels of health literacy exhibit lower engagement and adherence to medical and psychological treatment protocols (Maura & de Mamani, 2017; Raue & Sirey, 2011; Thapa & Nielsen, 2021). AI tools could help reverse this disparity by identifying factors associated with optimal treatment participation. But again, developers must continually assess, mitigate, and monitor AI models to ensure engagement and adherence are enhanced for all people.

Maintenance and monitoring.

After successful treatment of a psychological problem, the next goal is to maintain treatment gains over time and prevent relapse (Marlatt et al., 2009; Newman, 2012). This work may include helping the person transition from therapy successfully and helping them develop a plan about how to maintain treatment gains independently. Additional goals might include training individuals to self-monitor, recognize when symptoms are returning, and formulate a plan to reach out to obtain further treatment when needed. Although these strategies are important and effective components of the therapy termination process (Kupers, 1988), people may find it difficult to identify when relapse occurs or to recognize symptom recurrence early when relapse is easier to correct. AI could be used to help therapists identify early signs of relapse. People could provide data remotely (e.g., completing brief surveys at regular intervals submitted via smartphones or providing passive sensing data measuring sleep or exercise on wearables) so that algorithms can alert therapists when someone’s functioning begins to decline. Such algorithms may be particularly effective in identifying pre-relapse predictors of later functional decline (e.g., decreased physical activity, fewer social interactions). Relatedly, AI could be used to send micro-interventions or “nudges” at early stages in the relapse process (e.g., remember to go for a walk today), which could prevent a full relapse episode and the necessity for re-entering treatment. This application of AI could be especially helpful in bridging the gap between in-person therapy and treatment termination, allowing people to transition from therapy and gain skills independently while providing a safety net to monitor and quickly identify when additional support is needed. Some research indicates that people from marginalized groups are more likely to experience relapses of certain mental health problems; for example, racial and ethnic minoritized groups and those from under-resourced communities are disproportionately affected by poor health literacy, which has been shown to predict smoking relapse (Hoover et al., 2015; Stewart et al., 2014). AI may help to address such health disparities; however, as with all AI applications, scientists must carefully consider how marginalized and minoritized communities will be impacted.

Prediction.

Another important goal in the behavioral health field is to predict future mental health problems, episodes, or crises. Predicting who might need future mental health care provides an opportunity to administer preventative care, which is often less expensive and more effective than treating a mental health problem that has already developed (Gruber et al., 2021; Lawrie et al., 2019). Furthermore, accurate prediction helps clinicians and health policy administrators allocate resources efficiently and ensures that people receive optimal support. For example, AI could predict which pre-teens are at risk for developing depression. If effective, allocating resources toward these pre-teens could prevent them from developing mental health problems later in life. As with other mental health AI applications, decisions like this have important and potentially life-changing implications for the user. Inaccurate predictions could mean that individuals in need of services do not receive them or that people receive services they do not need. When algorithmic decisions are systematically biased against certain groups, those AI models will perpetuate rather than mitigate existing disparities in mental health care.

Inherent systemic social inequities result in biased foundations upon which we sample our data, select our teams, and build and evaluate our algorithms. Biased data and model development practices will result in biased algorithms, leading to biased applications. When people who share a certain aspect of social identity (e.g., race, age, gender) experience systematic bias caused by an algorithm's data and decision, the social inequities of the disadvantaged groups are fed back into the algorithm development system. This kind of bias is especially problematic in reinforcement learning models where the AI continually updates the model based on system feedback, causing incorrect or biased decisions to amplify over time.

An Emerging Field of Research in Engineering: "Fair-Aware AI"

With the rise of AI, including applications from business to health care, government, and more, AI developers are becoming increasingly aware that the systems they are building do not always reflect their values. Specifically, engineers have found that AI algorithms deployed in various contexts (e.g., job hiring, college admissions) are generating decisions biased against certain genders, races, ages, and ethnicities—even when they are not intended to or when the application appears, superficially at least, unrelated to the bias generated. Many of these examples have been highly publicized and have important social implications. For example, Buolamwini and Gebru (2018) found gender and racial bias in the performance of facial recognition algorithms that are widely deployed by Facebook, IBM, and Microsoft. The authors found that these algorithms are substantially better at classifying gender when looking at people with lighter compared to darker skin tones. Additionally, research shows that speech detection software commonly employed by smartphones and utilized in household smart devices generates more word recognition errors in women and people from minoritized backgrounds (Bajorek, 2019; Koenecke et al., 2020). Models that flag and remove hate speech on Twitter and other online platforms have inaccurately flagged African Americans' tweets as hate speech and systematically over-removed tweets posted by African American people (Davidson et al., 2019; Sap et al., 2019). Similarly, LBGTQIA+ YouTubers, whose careers and livelihood rely on the distribution and promotion of their content online, have documented that YouTube algorithms systematically over-flag and over-remove LBGTQIA+ content, leading them to file a discrimination lawsuit against the company (Bensinger & Albergotti, 2019). In yet another example, a recidivism prediction software called COMPAS has been demonstrated to overestimate the recidivism risk of Black and African American offenders compared to White offenders (Angwin et al., 2016; Dressel & Farid, 2018). Predictive policing systems have similarly come under scrutiny for their lack of transparency and training programs based on biased data that could lead to discrimination against minoritized communities (Akpinar et al., 2021; Richardson et al., 2019).

Examples of systematic AI bias in the health care domain are also accumulating. For example, Obermeyer et al. (2019) explored racial bias in risk prediction algorithms used by the health care insurance industry to refer people for additional screening procedures. The optimization metric used to define the algorithm was health care spending, which was used as a proxy for health, even though health care spending and health are not equivalent. Due to disparities in health care access and overall public health awareness for racial and ethnic minoritized groups, less money is spent on their care relative to non-Hispanic White patients. Thus, Black and African American patients had to have more severe and impairing symptoms than non-Hispanic White patients on average to reach the same health care spending level and to receive the same health risk score. This also meant that Black and African American patients had to be sicker on average before they were referred for additional services, thereby perpetuating the pre-existing inequity upon which the algorithm was built. In another example, Kamulegeya et al. (2019) found that an algorithm used to diagnose skin conditions built on a primarily non-Hispanic White sample performed far below acceptable levels when applied to people with other skin tones. Such low levels of model performance could lead to the misdiagnosis or underdiagnosis of medical conditions for minoritized groups and further perpetuate health disparities. To date, documented cases of AI-related bias in mental health applications are less common than in health care generally. But, as AI-based mental health models are increasingly built and deployed, instances of bias in the mental health domain will likely increase (Rubeis, 2022). Importantly, the conceptualization and expression of emotion and mental health vary considerably across gender, race, ethnicity, and culture (Hareli et al., 2015; Office of the Surgeon General, 2001). Many AI applications for mental health aim to identify emotional states using Natural Language Processing (NLP) technologies, including the identification, transcription, and analysis of speech and text data (e.g., Calvo et al., 2017; Le Glaz et al., 2021); however, developers should exercise caution, as NLP has been shown to produce significant performance biases across different genders, races, ethnicities, religions, sexualities, cultures, and ages (Straw & Callison-Burch, 2020). Outside of the mental health domain specifically, psychological scientists are increasingly recognizing the risk of bias in AI in research generally, as evidenced by one recent article by Tay and colleagues (2022) that outlines measures for mitigating AI-based measurement bias in psychological assessment.

Growing recognition of how AI-related bias perpetuates societal inequities has led to the emergence of a specialized area of research termed fair-aware AI (Feuerriegel et al., 2020; Jones et al., 2020; Mehrabi et al., 2021; Veale et al., 2018). This small but fast-growing field of study focuses on developing methodologies and standards for building values into AI systems and identifying and mitigating AI-related bias. These ideas, which emerged in the engineering field, are now spreading to applied fields, including medicine, although applications in mental health have not yet been widely implemented. In the following sections, we summarize current methods for assessing and mitigating bias in AI and recommendations for how these methods can be integrated and employed for mental health application. Our review of these methods includes (1) a summary of the types of AI bias and the junctures in the algorithm development process where bias can be introduced, (2) a summary of the current methods for assessing and mitigating bias in AI, and (3) recommendations for psychological scientists to integrate fair-aware standards in AI mental health applications.

Types of AI Bias and Junctures Where Bias May Infiltrate the Model Development Process

To reduce bias in AI applications for mental health, it is important to first identify the types of bias that can occur and the points at which it is likely to be introduced in the model development process. An AI-powered system's decisions reflect the type of input data used and decisions made during training and evaluation; therefore, if the datasets and processes imbibed by an AI model are discriminatory (and adequate corrections are not employed to fully remove the bias), the output recommendations will also be discriminatory. Bias can be introduced at any point in the development process, including its sociocultural foundations, data input, model building, performance evaluation, and deployment. Bias may be introduced due to negligence, lack of understanding about the implications of the bias, or because the data scientists working on the training process are prejudicial. Figure 1 provides an overview of the algorithm development process and points where bias may be introduced. Table 1 summarizes different types of biases that could harm the predictions of a model at each phase of the algorithm development process. For a comprehensive review of the types of bias in AI, see Mehrabi et al. (2021).

The **sociocultural foundation** stage generally refers to society's systematic and historically biased foundations. Sociocultural foundation bias impacts all subsequent steps in the model-building process. An example of bias introduced at this stage includes *structural inequality bias,* which occurs when biases woven into the fabric of institutions, government, and organizations cause one group of people to be favored over another. For example, children without access to high-quality educational resources may have lower test scores; individuals receiving approval for a bank loan may be more likely to start a business or own a home; people with a history of incarceration are at higher risk for re-incarceration. If we aim to build algorithms to predict important outcomes, then an important foundational step would be to consider how structural inequalities might relate to the outcome and thus lead to biased results. Similarly, *historical bias* occurs when AI models are trained on data that no longer accurately reflect current reality (e.g., shifting trends in the gender pay gap; Australian Human Rights Commission, 2020). Models built on biased foundations caused by structural inequity or historical bias are often biased; these foundational biases must be considered before beginning the development process.

The **data collection** stage refers to collecting data used to build the model. The data collected will reflect the biases of the sample from which it was sampled. An example of bias introduced during the data collection stage includes *representation bias,* where certain groups are inadequately represented in the sample. Most models optimize results for average accuracy, which means that the majority demographic (and the associated estimates of parameters linking features to labels) receives the most weight in the algorithm construction. If the association between the input features and the labels differs across groups, the majority group will be more accurately represented. *Measurement bias* occurs when features used in algorithm development fail to accurately represent the group for which the algorithm was developed. For example, measurement tools for capturing emotion in one culture may not accurately represent the experience and expression of emotion in different cultures. Or in ambulatory monitoring, measurement bias can occur when people from different groups use devices or equipment of differing quality to record their data.

The **model-building** stage refers to the process of decision-making when building the AI model. For instance, *label bias* refers to bias that occurs when selecting the criteria used to label the predicted outcomes. Systematic bias in categorizing labels can also lead to discrimination against certain groups. In the example of hate speech classifications on Twitter, the initial labels generated to build the model failed to accurately account for language differences across people from different racial and ethnic backgrounds; the unaccounted differences in language resulted in the tweets of African American individuals being inaccurately flagged as abusive speech (Davidson et al., 2019; Sap et al., 2019).

The **model performance** stage refers to biases that emerge when biases in the preceding stages have deleterious effects on the evaluation and performance of the model. Biases in model performance are typically caused by *class imbalance bias,* where there are few examples from historically underrepresented groups in the dataset compared to many examples of the majority group. This bias occurs when a model performs more accurately for some groups compared to others or results in a higher false positive or false negative rate relative to the dominant group.

Finally, the **human inference and deployment** stage refers to bias during algorithm deployment that exacerbates the inherent bias not previously addressed. In *feedback loop bias,* biased algorithms can lead to biased decision making, which further disadvantages marginalized groups, exacerbates social and systemic foundational biases, and feeds into subsequent model development.

Assessing and Mitigating Bias in AI

In the following sections, we review current methods for assessing and mitigating bias in AI. The suggestions provided, summarized from the literature to date, are not an exhaustive list. As a relatively new field, standards for fair-aware AI are still evolving; however, the evidence validating the use of these methods is growing rapidly and has already accumulated an impressive empirical base (e.g., Feng, 2022; Kamishima et al., 2012; Mehrabi et al., 2022; Oneto et al., 2019; Pfohl et al., 2021; Ustun, 2019; Zemel et al., 2013; Zhao et al., 2018). Although a complete review of all possible mitigations methods is outside the scope of the paper, we provide a general overview here to guide development and implementation within psychology. As fair-aware grows into a specialized subdomain of AI, we can expect many new developments in this area of research. The following recommendations are a starting place for mental health researchers to begin incorporating bias mitigation strategies into the algorithms they develop for mental health. In general, we propose that as standard procedure when developing algorithms for mental health applications, researchers conduct systematic bias assessments and employ mitigation strategies where needed. We further propose that researchers publish methods and results of bias assessments and mitigation strategies and make this information available to the public before implementing algorithms in research, health, or commercial settings.

Assessment criteria.

Evaluation of bias in AI algorithms should include three general criteria: appropriateness, statistical bias, and fairness (Fletcher et al., 2021). Before beginning development, researchers should determine whether the task is *appropriate* for an AI. For example, why is it preferable for an AI, rather than a human decision-maker, to execute the task? What are the potential benefits and risks of using an AI? Is the proposed AI appropriately matched to its intended context? Developers must consider these fundamental questions before proceeding to the next assessment phase. After the initial model has been developed, researchers should consider whether the algorithm exhibits *statistical bias* concerning *sensitive class*. A *sensitive class* refers to any shared identities or backgrounds that could result in discrimination against certain people or groups and include: race, ethnicity, color, religion, national origin, sex, pregnancy, sexual orientation, gender identity, age, body size, physical or mental disability, veteran status, genetic information, citizenship status, culture, or spoken language, among others. The term *statistical bias* refers to the mathematical result of an algorithm and occurs when an algorithm exhibits a systematic error or an unexpected tendency to favor one outcome over another concerning a sensitive class (Mehrabi et al., 2021). Statistical bias is typically introduced by training the AI on an unrepresentative sample where the majority group obtains greater weight in determining the parameters estimates than the minority group due to be being better represented in the data.

Bias in algorithms often leads to *unfairness,* which occurs when a person or group systematically experiences more unfavorable outcomes than the majority group (Fletcher et al. 2021). While *statistical bias* refers to the mathematical properties of the AI, *fairness* refers to the ethical and legal implications resulting from the statistical bias. Fairness can be assessed in various ways but generally refers to how well an algorithm performs in each group, which when applied to important decisions regarding diagnosis and treatment, could have serious implications for the health and functioning of its users. *Equality of odds* is achieved when the positive error and negative error rates are equal across groups. Related fairness concepts include *equal outcomes*, which is achieved when the outcomes of the algorithm (e.g., a diagnosis made or treatment recommended) are equal across groups, and *equality of opportunity,* which is achieved when positive outcomes (i.e., true positive rates) are equal across groups.

In many applications, researchers developing algorithms should remove bias related to sensitive classes. However, it is also important to note that, in some instances, differences across groups may be expected or warranted. For instance, disorder rates may legitimately differ across groups (e.g., higher rates of sickle cell disease in Black and African American people). In these cases, the goal is not for groups to be diagnosed at equal rates but rather that the true proportions of the disorder are maintained across groups. That is, the ratio of diagnostic labels used to build the algorithms (the “ground truth”) should match the ratios of predictions generated by the algorithm. It is critical in these instances that decisions reflect true group differences rather than underlying bias. Bias in diagnosing a condition in an underrepresented group might lead a researcher to erroneously believe that the biased proportion represents a true group difference when the observed differences reflect systematic inequities. Often, determining what constitutes “truth” is a decision fraught with uncertainty because it is also subject to bias. This fact is especially salient in the mental health field, which focuses on the study and classification of internal and subjective experience, underscoring the importance of assigning labels carefully with input from the population upon which the algorithm will be deployed.

To date, various methods have been developed to reduce bias and ensure fairness in the design of AI models. These approaches, categorized into three stages of the development process, include data pre-processing, model in-processing, and decision post-processing (Mehrabi et al., 2021). The sections that follow outline steps to be completed at each stage. Figure 2 provides a flow chart of the decision-making process for completing a bias assessment and mitigation.

Data pre-processing.

Data pre-processing techniques, applied after the data are collected but before the model has been developed, aim to transform features and labels so that underlying differences across groups are removed. Specifically, after data are collected, researchers conduct statistical tests (e.g., ANOVA, *t*-test) to determine if any of the input features or the labels used to build the model show differences between groups or across the intersection of groups, (e.g., race × gender). Bias can occur in the input features when variables used to predict the outcome of interest vary systematically between different classes. In mental health applications, some models might use speech transcriptions as input data in an algorithm designed to detect mood states. However, AI-based speech recognition software is known to generate higher error rates in women and people from racial and ethnic minoritized backgrounds compared to other groups (Bajorek, 2019; Koenecke et al., 2020). If the data fed into the AI model better represent some segments of society compared to others (especially if these inputs are based on algorithmic decisions that exhibit bias), the resulting model could be biased as well. Similarly, bias in labels may occur if the criterion the algorithm is predicting is biased. As an example, bias could occur when people tasked with labeling conflict episodes in audio recordings misidentify conflict in individuals whose background differs from their own without taking into account how the expression of emotion and its interpretation vary across cultures (e.g., Hareli et al., 2015; Office of the Surgeon General, 2001). Because biased input features and labels can result in biased algorithms, scientists should examine whether these variables have systemic biases before moving on to the model-building phase.

Again, it is worth noting that, depending on the nature of the observed difference, a statistical difference across groups in features or labels may not be problematic. Observed differences across groups may be classified into three scenarios (1) differences that truly exist and should be maintained (e.g., vocal pitch is higher in women than in men due to biological factors); (2) differences that are likely the result of social biases (e.g., women speak fewer words than men; a group underperforms on an achievement test; a group has higher rates of recidivism) *or* the difference, regardless of its cause, could disadvantage a group *in the future* if maintained by the algorithm (i.e., a legacy effect); and (3), differences that are not true difference but which are the result of biased thinking or measurement (e.g., one group is over-diagnosed with a mental health problem due to inaccurate perceptions of outgroup behavior). Crucially, classifying any of the differences mentioned above is also a subjective decision that is vulnerable to bias. It is therefore imperative to assemble diverse model-building teams, engage in thorough stakeholder and target population interviews, and carefully consider the meaning and implications of observed differences. On the one hand, a diagnosis can help ensure that people needing resources are identified and supported. On the other hand, receiving a diagnosis may lead to stereotyping and stigmatization. We suggest that differences across groups in features or labels should be maintained in Scenario 1 but removed or corrected in Scenarios 2 and 3 because these scenarios likely reflect underlying social biases and maintaining them could cause harm. Additionally, after the algorithm has been deployed, it is crucial to conduct regular assessments of the algorithm's impact with input from the stakeholders and target populations to protect against unintended consequences, ensure the AI continues to work across developing and shifting social contexts, and allow for algorithm updates based on community feedback.

Once the development team has completed tests of group differences and stakeholder and target population interviews, they can take corrective action to remove biases from the dataset. For features, this might include removing the variable from the dataset or implementing a statistical correction, such as a data transformation or the recovery of and noise injection to inputs that correspond to sensitive attributes (Calmon et al., 2017; Zhang et al., 2018). Taking such action could eliminate sensitive attributes in the recorded signals, which might be contributing to the propagation of bias in AI decisions. For biased labels, a researcher could relabel the data to ensure an equal proportion of positive predictions for the sensitive group and its counterparts (Hardt et al., 2016; Luong et al., 2011). Other techniques for improving problematic models include reweighing labels before training (Feldman et al., 2015; Kamiran & Calders, 2012; Luong et al., 2011) and controlling target labels via a latent output (Kehrenberg et al., 2020). It is critical that members of sensitive classes provide feedback on assigned labels, especially when those labels are subjective and culturally grounded (e.g., coding tweets for abusive language or labeling audio recordings for episodes of conflict).

Model in-processing.

Model in-processing techniques attempt to modify the algorithm's training process to ensure fair treatment for all samples. One example of a model in-processing technique is fairness regularization. The fairness regularization technique gives higher importance to samples from the sensitive group (Gorrostieta et al., 2019; Kamishima et al., 2011), therefore promoting pattern learning specific to that group. Researchers have designed other objective functions to promote fairness, including penalizing the mutual information between the sensitive feature and the classifier predictions (Kamishima et al., 2012) and adding constraints to the loss functions that require satisfying a proxy for equalized odds or disparate impact (Woodworth et al., 2017; Zafar et al., 2017a, 2017b). A minimax optimization problem has also been formulated to enhance fairness in addition to regression and classification loss criteria (Berk et al., 2017). Privileged learning is another strategy that makes sensitive information available at training time but not at the testing time (Quadrianto & Sharmanska, 2017).

Decision post-processing.

Decision post-processing modifies the outcomes of the AI model to ensure that decisions are similar and accurate across groups. Once the algorithm has been developed and adequate model performance is achieved, researchers should examine systematic differences in model performance through statistical testing, such as *t*-tests. Specifically, researchers should examine the outcomes, or the ratio of model predictions (e.g., the number of diagnoses made versus not made), by group and by the intersection of groups (e.g., race × gender). The ratios should be the same across groups or, in the case of scenario 1 described above and assuming that the labels of the data are not biased, the AI algorithm should maintain the ratio of the labels (e.g., if there were 10 diagnoses and 20 non-diagnoses in the labels, the algorithm should predict a similar ratio). Researchers should also ensure that the proportion of negative and positive outcomes or likelihood of positive outcomes is the same for both groups (Hardt et al., 2016; Pleiss et al., 2017), as well as ensuring that favorable outcomes do not solely occur for the majority group (Kamiran et al., 2012).

Additionally, model performance metrics (e.g., accuracy, kappa, F1 score) should be compared across different groups. Models that perform significantly better in one group could lead to social inequities by benefitting certain groups more than others. Although related, model *outcomes* and model *performance* are distinct and should be examined separately. A model could have similar ratios of outcomes across groups (or maintain desired ratios across groups) but be far less accurate in one group than another or vice versa. The ratios of detected versus non-detected states may or may not differ across groups, depending on the application; however, a model should work equally well across groups in general. If unexpected differences in ratios of decision or model performance occur, then the researcher should either go back to the model-building stage and apply model-in-processing techniques or should apply decision post-processing methods as described above. This process should then be repeated until fairness is achieved. Examples of decision post-processing methods include taking a subset of samples and changing their predicted labels to appropriately meet a group fairness requirement (i.e., statistical measures of performance being similar across groups; Kamiran et al., 2012; Pleiss et al., 2017) or retrospectively making sure that the decision of the AI algorithm is the same for individuals that share similar characteristics, thus optimizing for individual fairness criteria (Lohia et al., 2019; Petersen et al., 2021; Yeom & Fredrikson, 2020). An important advantage of decision post-processing algorithms, compared to data pre-processing and model in-processing, is that they can be more computationally efficient because it does not require transforming the entire dataset or re-training the model at a large scale.

Integrated Equality- and Equity-Based Approaches to Mitigating Bias in AI

When implementing bias mitigation techniques, it can be helpful to think about these strategies in terms of equality-based and equity-based approaches for addressing health disparities. Equality-based approaches focus on ensuring that all groups have equal access to resources and foundations for pursuing achievement and health (e.g., supplying all children with high-quality education or access to preventative health care; Murphy & Hallinger, 1989; Oliver & Mossialos, 2004). Equity-based approaches focus on ensuring that people who come from different backgrounds with varying degrees of resources and opportunity achieve positive outcomes (e.g., hiring a diverse team to work at a company or admitting a diverse body of students to an educational program; Creary et al., 2021; Sturm, 2006). Figure 3 provides examples of bias in AI mitigation strategies falling within each category. For example, equality-based approaches might include ensuring that the development team is representative and that different groups are adequately represented in the data. Equity-based approaches might include applying model in-processing or decision post-processing methods to ensure fair outcomes of algorithms. As a method for mitigating health disparities, equality can be conceptualized as a bottom-up approach, where the source of the disparity is addressed. Equity, in contrast, can be conceptualized as a top-down approach, where existing disparities are corrected. We suggest that algorithm developers apply both equity- and equality-derived methods for an integrated and justice-oriented approach to fair-aware AI. Although we should always take steps to address the causes of biases in our data, models, and society at large, we must also recognize that we cannot instantaneously undo these systemic and endemic historical and foundational legacies. Equality-based AI approaches will lead to increased fairness, whereas equity-based AI approaches will mitigate the foundational inequalities that cause disparities in outcomes. When employed in tandem, equality- and equity-based approaches will contribute to more just outcomes over time.

Recommendations

The following sections summarize our recommendations for assessing and mitigating bias in AI applications for mental health. These recommendations are divided into four stages, including model building (selecting a team, collecting data, and building the model), model evaluation (determining how well the model works and whether it exhibits bias), bias mitigation (fixing observed biases), and model implementation (publishing, monitoring, and updating). See Table 2 for a list of these recommendations. We suggest that these standards are implemented at multiple levels, including by individual research groups, ethical boards that provide oversight, and as legal requirements.

Model building.

The first step of building fair-aware AI models is to recruit a diverse team of people to build and inform the development of the algorithm. Some early studies have indicated that a lack of diversity in development teams could adversely impact model results. For example, one study found that prediction errors were correlated within team members' demographic groups, especially for gender. That is, male programmers prediction errors were correlated more stronger to other males as compared to females. The study found that averaging across team members' gender introduced equally predictive but uncorrelated information, which created performance improvements through cross-demographic averaging (Cowgill et al., 2020). Team members should include the scientists and engineers who will build the algorithms, as well stakeholders (e.g., doctors, therapists, insurers) and community members (i.e., potential users of the technology) who represent the population for whom the algorithm is being built. Scientists, engineers, stakeholders, and community members should hold meetings at each stage in the development process so that members' feedback, critiques, revisions, and contestations can be effectively integrated and to minimize mistakes that may limit the future effectiveness, sensitivity, inclusiveness, fairness, and generalizability of the algorithm to be developed. Major public policy organizations and research institutions, including the Brookings Institute and the Center for Equity, Gender, and Leadership at the Hass School of Business at the University of California at Berkeley have proposed stakeholder engagement to help developers select inputs and outputs of certain automated decisions (e.g., Lee et al., 2019, Smith & Rustagi, 2020). Getting users engaged early and throughout the algorithm development process will ultimately improve the user experience. Before starting development, methods for obtaining feedback and the timings of the feedback sessions should be planned. We suggest holding feedback sessions at the following stages: (1) conceptualization of the problem, (2) choice of labels, data streams, and sample to be collected, (3) pre-processing assessment, (4) post-processing assessment, (5) pre-implementation, and (6) post-implementation monitoring. Questions administered at the meetings should be standardized and should thoroughly assess all team members' perceptions and concerns surrounding fairness, accountability, privacy, and data security. Mulligan et al. (2019) provides a helpful framework for selecting topics and designing questions to be covered in these meetings.

In the first stage, it is critical to interview the target population to ensure appropriate and accurate conceptualization of the problem and the solution. Labels and features used to build the model should be carefully selected with input and feedback from members of the target population. Avoid using sensitive attributes as feature inputs in model development as training AIs based on these proxies may unintentionally yield biases against specific groups. It is possible that algorithms may sometimes infer sensitive attributes via other variables that are non-sensitive themselves but related to the sensitive attribute. To avoid possible discriminatory outcomes, researchers are employing a myriad of techniques, including model balancing and encrypting sensitive attributes, which achieve accurate prediction performance while concurrently improving fairness (Kilbertus et al., 2018; Yan et al., 2020). Various discrimination-aware deep learning methods currently employ loss functions or maximization adaption networks to improve model performance (Serna et al., 2020; Wang & Deng, 2020; Wang et al., 2019). One important future direction in the fair-aware field includes creating empirical tests of whether an algorithm can unintentionally recover information about a sensitive group.

In the second stage, ensure that your sample includes people from diverse backgrounds and matches the population for whom the algorithm will be deployed. For example, diverse databases have been proposed to improve face biometrics in a multitude of studies (e.g., Buloamwini & Gebru, 2018; Hupont & Fernández, 2019; Kärkkäinen & Joo, 2019; Merler et al., 2019). In one example, Kärkkäinen and Joo (2019) created a novel face image dataset containing 108,501 images, with an emphasis on balanced race composition. They found the model trained on their dataset to be substantially more accurate on novel datasets and with consistent accuracy between race and gender groups. Merler et al. (2019) have also contributed to increased accuracy through their Diversity in Faces (DiF) dataset, which includes 1 million annotated human face images publicly available for advancing the study of facial diversity. Once data are collected, perform the pre-processing assessment, elicit feedback, and revise development plans based on the feedback received (Stage 3). Various data preprocessing methods have been shown to mitigate bias by selecting a subset of data that satisfies specified fairness constraints (e.g., representation rate) without attempting to model the distribution (Celis et al., 2016; 2018). In one applied example, Zhang et al. (2018) demonstrated that data pre-processing, including removing unwanted attributes from the signals, helps in mitigating biases when predicting income. If possible, use explainable and intuitive models and user-friendly visual tools so that non-specialists can easily understand the algorithm's inner working and accurately judge its social implications. Transparency in model development, variables used, and any underlying assumptions at all stages is critical to developing fair-aware algorithms.

Model evaluation.

After obtaining model results, researchers should follow a standard set of procedures to assess bias in the resulting algorithm (Stage 4). This includes examining the ratios of algorithmic outcomes (e.g., diagnosis versus non-diagnosis) for each sensitive class and the intersection of sensitive classes (e.g., race × gender). In the case of Scenario 1 outlined above, when there are true differences across groups, those differences should be maintained; thus, the ratios of the algorithmic outcomes should match the ratios observed in the labels. In the case of Scenario 2, where the difference might exist but should not be maintained, or Scenario 3, where the difference does not really exist but is the result of biased thinking and measurement, then bias mitigation strategies should be used to remove differences in the ratios of algorithmic outcomes across sensitive classes. Further, researchers should examine whether model performance metrics (e.g., F1 score, kappa, false, false positive rates, false negative rates) differ across classes. Ideally, models should perform equally well for all people and meet minimum performance levels in all groups. Finally, researchers should share the results of the post-processing assessment with stakeholders and community groups for feedback and revision. To increase transparency in the assessment of AI bias, researchers should share de-identified datasets and code in open data repositories.

Bias mitigation.

If bias is observed in the AI model as described above, then the researcher should return to the model development stage and apply model in-processing or post-processing bias mitigation strategies (Berk et al., 2017; Gorrostieta et al., 2019; Kamishima et al., 2011; Woodworth et al., 2017; Zafar et al., 2017a, 2017b). For example, researchers could transform data, inject or recover noise, (Calmon et al., 2017; Zhang et al., 2018), relabel the data to ensure an equal proportion of positive predictions for the sensitive group and its counterparts (Hardt et al., 2016; Luong et al., 2011), reweigh labels before training (Feldman et al., 2015; Kamiran & Calders, 2012; Luong et al., 2011), control target labels via a latent output (Kehrenberg et al., 2020), apply fairness regularization, penalize the mutual information between the sensitive feature and the classifier predictions (Kamishima et al., 2012), or add constraints to the loss functions that require satisfying a proxy for equalized odds or disparate impact (Woodworth et al., 2017; Zafar et al., 2017a, 2017b). See Figure 2 for a general framework for applying bias mitigation techniques. After applying corrections, researchers should repeat the previously delineated steps in the model-building phase until the model no longer yields biased results.

Although research in the fair-aware AI domain is still evolving, early and promising work shows that bias evaluation and mitigation methods are effective in reducing bias in AI algorithms (e.g., Feng, 2022; Hardt et al., 2016; Kamishima et al., 2012; Mehrabi et al., 2022; Oneto et al., 2019; Pfohl et al., 2021; Ustun, 2019; Zemel et al., 2013; Zhao et al., 2018). For example, Ustun et al. (2019) used a recursive procedure to decouple subgroups of individuals and train decoupled classifiers for each subgroup. The authors showed that this process improved estimation accuracy while ensuring fair outcomes across the identified subgroups. In another example, Oneto et al. (2019) imposed fairness constraints, measured via Equalized Odds or Equal Opportunities, on a multitask learning framework that clustered individuals based on their demographic attributes, as estimated by the input data. These were applied for the prediction of violent recidivism scores, with results indicating high accuracy and fairness across groups. Kamishima et al. (2012) proposed a “prejudice remover regularizer,” which enforces the outcome of the machine learning model to be independent of sensitive information. This was applied to remove bias to a machine learning task that estimates high incomes based on other characteristics, finding that this approach successfully reduced gender and racial bias in the model. Pfohl et al. (2021) attempted to characterize the impact of imposing group fairness constraints on model performance for clinical risk prediction. Results showed that the method reduced bias while degrading the overall performance of risk prediction, highlighting the need for developing methods that sustain fairness for clinical applications. Hardt and colleagues (2016) showed that using different metrics to train and evaluate their model yielded fairer decisions for predicting credit scores. In a similar example, Zemel et al. (2013) added fairness constraints during training and found that this method mitigated age bias when predicting credit scores. Finally, Zhao et al. (2018) imposed post-processing to remove gender bias in word embeddings, resulting in language representations with enhanced fairness between men and women.

Model implementation.

Upon building a fair-aware algorithm that follows the guidelines listed above, developers should conduct interviews and qualitative assessments with the target population and other stakeholders to evaluate the model’s potential impacts on the community (Stage 5). Assessment should also include the target group’s perceptions of the algorithm’s fairness and concerns relating to privacy, security, and accountability. Before dissemination, the team should write an implementation plan that includes a list of possible negative outcomes and worst-case scenarios, planned remedial actions to be taken if such events occur, safeguards for preventing algorithmic misuse, a monitoring plan, defined parameters for evaluating the algorithm, and stopping rules for discontinuing the algorithm’s use, if warranted. If possible, these plans should include an opt-out option and an appeal process to request human review of decisions generated by the algorithm. Such appeal processes should be efficient and easily accessible to users. Reasonable alternatives to the AI should be offered and should, to the extent possible, be comparable in their cost and effectiveness. De-identified data, model results, the bias assessment, and implementation plan should be published and made publicly accessible. Additionally, published materials should be transparent and easily understandable to non-experts. After deployment (Stage 6), regular algorithm monitoring and calibration should be maintained, especially in the case of reinforcement models. Lastly, researchers should examine the interaction of the algorithm with the community members to evaluate potential algorithm drift and identify and correct unanticipated negative consequences, conduct regular assessments with stakeholders, incorporate revisions at regular intervals, and repeat the model development process if the algorithm is to be adapted to new populations or applications. Various AI advocacy groups, including a group of experts appointed by the European Commission (High-Level Expert Group on Artificial Intelligence, 2020) and the US Department of Commerce National Institute of Standards and Technology (Schwartz et al., 2020) have made similar recommendations in recent years. Although empirical work in this area is limited to date, one study by Shin (2021) showed that increased AI explainability was associated with increased levels of consumer trust and emotional confidence.

Importantly, the goal in designing AI applications for mental health is to build equitable and highly effective therapies. When ensuring equality of performance through equality of odds or opportunity, for example, it is possible to create a solution that works equally well for Groups A and B but does not work as well as it could for Group A because the performance of Group A was penalized in the loss function to ensure equity with Group B. Alternatively, it is possible that the therapy works poorly in both groups. The discipline of Distributive Justice involves determining the moral distribution of resources in society (Lamont & Favor, 2017). In the Distributive Justice framework, there are many competing philosophies, ranging from Strict Egalitarianism, which advocates for equal outcomes in all scenarios, to Libertarianism, which advocates that performance in the dominant group is never sacrificed, as that would constitute redistribution of resources at the expense of the dominant group’s liberty. Although we cannot provide definitive answers to these complex social issues, it is critical for researchers building AIs to consider how their applications intersect with these important philosophical, economic, social, and political questions. Some early work in the engineering domain has begun to build quantitative solutions for ensuring fairness while not sacrificing model performance (e.g., Obermeyer et al., 2019). For example, common optimization metrics used in machine learning, such as the disparate impact or equal opportunity difference, reflect the rationale behind these frameworks. However, after training the new unbiased models, it is essential to investigate the performance of the new algorithms for each considered group.

It is worth noting that it may not be possible or even desirable for researchers to ensure complete fairness for all sensitive classes and combinations thereof. Data on all sensitive class memberships are not always available and a complete assessment of all permutations of class memberships could prove unfeasible for some development teams. Problems relating to (1) inadequate power to test all group combinations, (2) differential power across groups given that groups may have differential representation in the data, and (3) inflated Type I error rates resulting from testing differences across many groups limit the ability to conduct comprehensive tests for all bias possibilities. Although the recommendations listed above outline the ideal scenarios for model development, we also recognize that not all teams will be able to complete all steps as recommended above for all class combinations.

At a minimum, algorithms should be developed and validated on the populations for whom they are intended, and the disseminated materials should include transparent information about the people and use cases for whom the output is unvalidated. Differences that remain after bias mitigation has been implemented should be clearly delineated in the published materials. Users of algorithms can then rely on this information to decide and consent to using the AI, but only if there is transparency in the model-building process (e.g., via interpretable or explainable AI algorithms) and comparable alternatives to using the AI are available and easily accessible.

A Call to Action: Standards of AI Bias Assessment and Mitigation in Psychological Science

Social, political, and historical events (Roberts & Rizzo, 2021; Williams, 2019) co-occurring with recent calls from researchers to dismantle systematic oppression within science broadly (Andoh, 2021; Buchanan & Wiklund, 2020; Byrd et al., 2021; Galán et al., 2021; Hochman & Suyemoto, 2020) have led to an increased focus on advancing diversity, equity, inclusion, and justice objectives in the mental health field. One central goal of psychological science in the coming years will be to develop mental health interventions that are accessible and effective for all people (Gee et al., 2021; Gruber et al., 2021; Ramos et al., 2021). Advances in technology and AI hold great promise for increasing the reach and impact of mental health care; this enthusiasm is evidenced by the amount of money invested, research grants awarded, training programs started, centers formed, and new positions created that focus on applying data science methods to mental health applications. However, the field must ensure that we do not repeat past mistakes or perpetuate past inequities and that the interventions we develop with these new technologies are widely accessible, equitably effective, and inclusively made. Although the concepts in this paper (e.g., collecting diverse samples, employing diverse teams) are broadly relevant to all research using inferential statistics, we contend that these ideas are especially important in the application of AI due to (1) the historical timing of the rising use of AI, (2) the great potential for future harm, as AI will soon create new systematized social structures at a scale that has been previously unseen, and (3), if heeded, an opportunity to prevent misapplication of AI in the mental health field. It is thus critical to generate awareness of what is at stake and develop methodological standards for assessing and mitigating bias in AI in our published research and mental health interventions deployed. The assessment and mitigation of bias in AI is quickly becoming a field of study unto itself and will help to ensure that the new technologies we develop are consistent with the values of our field to promote equity, inclusion, and fairness. In the coming years, as the field increases its focus on building data science applications for mental health, we must ensure that the next generation of scientists is trained and well-versed in applying fair-aware AI methodologies.

Acknowledgments

This project is based on work supported by NIMH Grant No. R42MH123368 (Timmons and Comer, Co-PIs), NSF GRFP Grant No. 1930019 (Timmons, PI), and NSF GRFP Grant No. DGE-0937362 (Timmons, PI). The contents of this paper are the responsibility of the authors and do not necessarily represent the official views of NIH or NSF. Adela C. Timmons and Matthew W. Ahle own intellectual property and stock in Colliga Apps Corporation and could benefit financially from the commercialization of related research.

Reflexivity Statement

A **reflexivity statement** is a statement that describes scientists' personal backgrounds to acknowledge how the outcomes of our research are influenced by and situated within specific biases, contexts, and lived experiences.

Adela C. Timmons: ACT is an Assistant Professor of Psychology whose research focuses on the intersection of data science and clinical psychological application. Through her research, she aims to build and apply novel technologies to increase the reach and impact of mental health care for people from historically under-resourced and under-served backgrounds. She is committed to understanding and developing strategies to mitigate disparities in mental health arising from big data and artificial intelligence. As a woman scientist specializing in data science and an entrepreneur, she aims to promote female achievement in fields where women's voices are under-represented and under-valued. ACT acknowledges the risk of bias stemming from her social identity as a White cis-gender female and aims to continually learn and expand her awareness of the impacts of oppression and systemic inequities on minoritized communities.

Jacqueline B. Duong: JBD is currently a clinical science doctoral student. Her research interest in digital mental health, with a specific focus on improving mental health outcomes for racial and ethnic minoritized groups and immigrant populations, stems from her own experience as a first-generation college student, cis gender woman, and daughter of Vietnamese boat refugees. Given her background, JBD is particularly motivated to build technologies that improve access to mental health care for under-represented and under-resourced groups.

Natalia Simo Fiallo: NSF has a bachelor's in psychology and is currently the research coordinator of a NIH-funded research study centered on families from under-resourced and minoritized backgrounds. One of her main interests includes expanding the accessibility to mental health care, especially for immigrants. As a Hispanic immigrant and first-generation college graduate, NSF is determined to contribute to clinical psychological science through awareness of biases that undermine effective treatment and psychoeducation of underserved populations.

Theodore Lee IV: TL earned his bachelor's in psychology and has developed a growing interest in using innovative technologies to enhance mental health care treatments. He is especially interested in improving the mental health trajectories of those who come from historically minoritized groups. His interest in this population draws on his own intersectionality as a Bisexual African American male. He understands the risk that bias poses to these digital mental health technologies. He will utilize his experience to mitigate the introduction of bias into the research and development of new digital mental health treatments so those from minoritized groups are not harmed.

Huong (Penny) Phuc Quynh Vo: PV is a computer engineering undergraduate student. She is interested in research related to the human mind, emotion, and action. She loves digging into messy data and being its translator to tell other people its messages and stories. Bias is something that cannot be removed; however, it can be mitigated. She wants to make the world a better place where everyone is treated equitably.

Matthew W. Ahle: MWA is a software developer and data scientist. He is the co-founder and CTO of Colliga Apps, a platform for digital mental health. His work background includes project management and IT strategy, web design and development, and UI/UX design. His current work at Colliga Apps involves the integration of data science algorithms with web and mobile applications. As an entrepreneur and software developer at the nexus of mental health care research and the tech industry, MWA is committed to building technological systems to enhance treatment outcomes and extend access to mental health care for people from traditionally under-represented and under-served backgrounds.

Jonathan S. Comer: JSC is a Professor of Psychology and parent who studies children’s mental health challenges and their treatment. Much of his work examines digital technologies and remote care formats and their potential for expanding the scope, responsiveness, and reach of children’s mental health services and supports. JSC recognizes that, despite early empirical support for technology-based care, research to date has often been constrained by non-representative sampling and digital mental health formats have yet to actualize their potential for meaningfully expanding treatment accessibility to new populations underserved by traditional brick-and-mortar services. He believes considerable efforts are needed to ensure digital mental health services go beyond simply adding more treatment options and alternatives for populations that are already served by traditional care models and that there is a critical need to understand how marginalized and minoritized populations experience and respond to digital mental health services. JSC recognizes that there is an inherent risk of bias associated with his position as a White cis-gender male studying how marginalized and minoritized populations experience technology-based care and the impact of such technology-based care on clinical outcomes.

LaPrincess C. Brewer: LCB is an Assistant Professor of Medicine in the Division of Preventive Cardiology within the Department of Cardiovascular Medicine at the Mayo Clinic in Rochester, Minnesota. She has a primary research focus in developing strategies to reduce and ultimately eliminate cardiovascular disease health disparities in racial and ethnic minority populations and in underserved communities through health promotion and community-based participatory research. LCB also has special interest in increasing minority and women’s participation in cardiovascular clinical trials through mobile health (mHealth) interventions. Additionally, she has published work on faith-based interventions for cardiovascular disease prevention, racial differences in weight maintenance, and psychosocial factors influencing cardiovascular risk factors.

Stacy L. Frazier: SLF is a Professor of Applied Social and Cultural Psychology. She directs a program of dissemination, implementation, and services research to support youth service systems in communities where structural racism and systemic injustice contribute to mental health disparities. Through community-engaged research and mixed method designs, she seeks to center and amplify the voices of providers in non-specialty settings serving Black and Latinx communities that are underserved in systems of care and underrepresented in psychological science. SLF acknowledges the risk of bias that accompanies her identity as a White cisgender woman and her position as a cultural outsider to the marginalized and minoritized communities with whom she collaborates. Her work builds on individual and cultural strengths, responds to the expressed needs, and respects the priorities, resources, and constraints of youth service systems.

Theodora Chaspari: TC is an Assistant Professor in Computer Science and Engineering. Her research interests lie in machine learning, data science, and affective computing. As a woman computer scientist and engineer, diversity, equity, inclusion, and justice issues are deeply important to her, rendering her close to the minoritized populations that are studied in this paper. TC believes that creating fair technologies will allow us to take a small step toward equitable mental health care.

References

Abe-Kim J, Takeuchi DT, Hong S, Zane N, Sue S, Spencer MS, Appel H, Nicdao E, & Alegría M (2007). Use of mental health–related services among immigrant and US-born Asian Americans: results from the National Latino and Asian American study. American Journal of Public Health, 97(1), 91–98. 10.2105/ajph.2006.098541 [PubMed: 17138905]

Akpinar N-J, De-Arteaga M, & Chouldechova A (2021). The effect of differential victim crime reporting on predictive policing systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 838–849). 10.1145/3442188.3445877

Alegría M, Nakash O, & NeMoyer A (2018). Increasing equity in access to mental health care: a critical first step in improving service quality. World Psychiatry, 17(1), 43–44. 10.1002/wps.20486 [PubMed: 29352534]

Amazon Web Services. Machine learning in the AWS cloud: Add intelligence to applications with Amazon SageMaker and Amazon Rekognition. Amazon https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias-metric-class-imbalance.html

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5 ed.). 10.1176/appi.books.9780890425596

American Psychiatric Association. (2017). Mental health disparities: Diverse populations. https://www.psychiatry.org/psychiatrists/cultural-competency/education/mental-health-facts

American Psychological Association. (2012). Recognition of psychotherapy effectiveness. Council Policy Manual. https://www.apa.org/about/policy/resolution-psychotherapy

American Psychological Association. (2020). Understanding psychotherapy and how it works. https://www.apa.org/topics/psychotherapy/understanding

Anderson-Lewis C, Darville G, Mercado RE, Howell S, & Di Maggio S (2018). mHealth technology use and implications in historically underserved and minority populations in the United States: Systematic literature review. JMIR mHealth and uHealth, 6(6), e128–e128. 10.2196/mhealth.8383 [PubMed: 29914860]

Andoh E (2021). Psychology’s urgent need to dismantle racism. Monitor on Psychology, 52(3). https://www.apa.org/monitor/2021/04/cover-dismantle-racism

Angwin J, Larson J, Mattu S, & Kirchner L (2016). Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. The American Psychologist, 61(4), 271–285. 10.1037/0003-066X.61.4.271 [PubMed: 16719673]

Arun V, P V, Krishna M, A BV, P SK, & S V (2018). A boosted machine learning approach for detection of depression. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 41–47). 10.1109/SSCI.2018.8628945

Australian Human Rights Commission. (2020). INFOGRAPHIC: Historical bias in AI systems. The Australian Human Rights Commission. Retrieved March 12 from https://humanrights.gov.au/about/news/media-releases/infographic-historical-bias-ai-systems

Bajorek JP (2019). Voice recognition still has significant race and gender biases. Harvard Business Review. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

Barello S, Triberti S, Graffigna G, Libreri C, Serino S, Hibbard J, & Riva G (2016). eHealth for patient engagement: A systematic review [Review]. Frontiers in Psychology, 6. 10.3389/fpsyg.2015.02013

Barlow DH, Farchione TJ, Bullis JR, Gallagher MW, Murray-Latin H, Sauer-Zavala S, Bentley KH, Thompson-Hollands J, Conklin LR, Boswell JF, Ametaj A, Carl JR, Boettcher HT, & Cassiello-Robbins C (2017). The unified protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: A randomized clinical trial. JAMA Psychiatry, 74(9), 875–884. 10.1001/jamapsychiatry.2017.2164 [PubMed: 28768327]

Battista P, Salvatore C, Berlingeri M, Cerasa A, & Castiglioni I (2020). Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. Neuroscience & Biobehavioral Reviews, 114, 211–228. 10.1016/j.neubiorev.2020.04.026 [PubMed: 32437744]

Belluz J (2019). Research fraud catalyzed the anti-vaccination movement. Let's not repeat history. How Andrew Wakefield's shoddy science fueled autism-vaccine fears that major studies keep debunking. Vox. https://www.vox.com/2018/2/27/17057990/andrew-wakefield-vaccines-autism-study

Benjamin LT Jr. (2005). A history of clinical psychology as a profession in America (and a glimpse at its future). Annual Review of Clinical Psychology, 1(1), 1–30. 10.1146/annurev.clinpsy.1.102803.143758

Bensinger G, & Albergotti R (2019). YouTube discriminates against LGBT content by unfairly culling it, suit alleges. The Washington Post. https://www.washingtonpost.com/technology/2019/08/14/youtube-discriminates-against-lgbt-content-by-unfairly-culling-it-suit-alleges/

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, & Roth A (2017). A convex framework for fair regression. arXiv preprint arXiv:1706.02409.

Blease C, Kharko A, Annoni M, Gaab J, & Locher C (2021). Machine learning in clinical psychology and psychotherapy education: A mixed methods pilot survey of postgraduate students at a Swiss University. Frontiers in Public Health, 9. 10.3389/fpubh.2021.623088

Bohr A, & Memarzadeh K (2020). The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare, 25–60. 10.1016/B978-0-12-818438-7.00002-2

Bravo C, O'Donoghue C, Kaplan CP, Luce J, & Ozanne E (2014). Can mHealth improve risk assessment in underserved populations? Acceptability of a breast health questionnaire app in ethnically diverse, older, low-income women. Journal of Health Disparities Research and Practice, 7(4), 6. https://pubmed.ncbi.nlm.nih.gov/25705576

Buchanan NT, & Wiklund LO (2020). Why clinical science must change or die: Integrating intersectionality and social justice. Women & Therapy, 43(3–4), 309–329. 10.1080/02703149.2020.1729470

Buolamwini J, & Gebru T (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v81/buolamwini18a.html

Byrd DA, Rivera Mindt MM, Clark US, Clarke Y, Thames AD, Gammada EZ, & Manly JJ (2021). Creating an antiracist psychology by addressing professional complicity in psychological assessment. Psychological Assessment, 33(3), 279–285. 10.1037/pas0000993 [PubMed: 33779204]

Cai H, Han J, Chen Y, Sha X, Wang Z, Hu B, Yang J, Feng L, Ding Z, Chen Y, & Gutknecht J (2018). A pervasive approach to EEG-based depression detection. Complexity, 2018, 5238028. 10.1155/2018/5238028

Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, & Varshney KR (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (995–4004 ed., pp. 995–4004).

Calvo RA, Milne DN, Hussain MS, & Christensen H (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649–685. 10.1017/S1351324916000383

Carbonell Á, Navarro-Pérez J-J, & Mestre M-V (2020). Challenges and barriers in mental healthcare systems and their impact on the family: A systematic integrative review. Health & Social Care in the Community, 28(5), 1366–1379. 10.1111/hsc.12968 [PubMed: 32115797]

Carpenter SM, Menictas M, Nahum-Shani I, Wetter DW, & Murphy SA (2020). Developments in mobile health just-in-time adaptive interventions for addiction science. Current Addiction Reports, 7(3), 280–290. 10.1007/s40429-020-00322-y [PubMed: 33747711]

Choi SB, Lee W, Yoon JH, Won JU, & Kim DW (2018). Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. Journal of Affective Disorders, 231, 8–14. 10.1016/j.jad.2018.01.019 [PubMed: 29408160]

Chouldechova A, & Roth A (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.

Christie RH, Abbas A, & Koesmahargyo V (2021). Technology for measuring and monitoring treatment compliance remotely. Journal of Parkinson's disease, 11(s1), S77–S81. 10.3233/JPD-212537

Churchwell K, Elkind MS, Benjamin RM, Carson AP, Chang EK, Lawrence W, Mills A, Odom TM, Rodriguez CJ, & Rodriguez F (2020). Call to action: Structural racism as a fundamental driver of health disparities: a presidential advisory from the American Heart Association. Circulation, 142(24), e454–e468. 10.1161/CIR.0000000000000936 [PubMed: 33170755]

Clough BA, & Casey LM (2011). Technological adjuncts to increase adherence to therapy: A review. Clinical Psychology Review, 31(5), 697–710. 10.1016/j.cpr.2011.03.006 [PubMed: 21497153]

Cooper A (2020). Academia's reluctance to market data science programs threatens global R&D. Big Data Made Simple. https://bigdata-madesimple.com/academia-reluctance-data-science-programs/

Creary SJ, Rothbard N, & Scruggs J (2021). Improving workplace culture through evidence-based diversity, equity and inclusion practices. 10.31234/osf.io/8zgt9

Cuijpers P, & Gentili C (2017). Psychological treatments are as effective as pharmacotherapies in the treatment of adult depression: A summary from Randomized Clinical Trials and neuroscience evidence. Research in Psychotherapy: Psychopathology, Process, and Outcome, 20(2). 10.4081/ripppo.2017.273

Davenport T, & Kalakota R (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94–98. 10.7861/futurehosp.6-2-94

Davidson T, Bhattacharya D, & Weber I (2019). Racial bias in hate speech and abusive language detection datasets. Proceedings of the Third Workshop on Abusive Language Online, 25–35. 10.18653/v1/W19-3504

De Houwer J (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. Perspectives on Psychological Science, 14(5), 835–840. 10.1177/1745691619855638 [PubMed: 31374177]

Deer B (2011). How the case against the MMR vaccine was fixed. BMJ, 342. 10.1136/bmj.c5347

Delgado-Rodriguez M, & Llorca J (2004). Bias. Journal of Epidemiology & Community Health, 58(8), 635–641. 10.1136/jech.2003.008466 [PubMed: 15252064]

Dixon LB, Holoshitz Y, & Nossel I (2016). Treatment engagement of individuals experiencing mental illness: review and update. World Psychiatry, 15(1), 13–20. 10.1002/wps.20306 [PubMed: 26833597]

Dobias ML, Sugarman MB, Mullarkey MC, & Schleider JL (2022). Predicting Mental Health Treatment Access Among Adolescents With Elevated Depressive Symptoms: Machine Learning Approaches. Administration and Policy in Mental Health and Mental Health Services Research, 49(1), 88–103. 10.1007/s10488-021-01146-2 [PubMed: 34213666]

Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, & Glozier N (2011). A systematic review of the impact of adherence on the effectiveness of e-therapies. Journal of Medical Internet Research, 13(3), e52. 10.2196/jmir.1772 [PubMed: 21821503]

Dressel J, & Farid H (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. 10.1126/sciadv.aao5580 [PubMed: 29376122]

Dwyer DB, Falkai P, & Koutsouleris N (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. Annu Rev Clin Psychol, 14, 91–118. 10.1146/annurev-clinpsy-032816-045037 [PubMed: 29401044]

Erguzel TT, Sayar GH, & Tarhan N (2016). Artificial intelligence approach to classify unipolar and bipolar depressive disorders. Neural Computing and Applications, 27(6), 1607–1616. 10.1007/s00521-015-1959-z

Evans GW (2004). The environment of childhood poverty. American Psychologist, 59(2), 77. 10.1037/0003-066X.59.2.77 [PubMed: 14992634]

Fadus MC, Ginsburg KR, Sobowale K, Halliday-Boykins CA, Bryant BE, Gray KM, & Squeglia LM (2020). Unconscious bias and the diagnosis of disruptive behavior disorders and ADHD in African American and Hispanic youth. Academic Psychiatry, 44(1), 95–102. 10.1007/s40596-019-01127-6 [PubMed: 31713075]

Faedda GL, Ohashi K, Hernandez M, McGreenery CE, Grant MC, Baroni A, Polcari A, & Teicher MH (2016). Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. Journal of Child Psychology and Psychiatry, 57(6), 706–716. 10.1111/jcpp.12520 [PubMed: 26799153]

Feczko E, Miranda-Dominguez O, Marr M, Graham AM, Nigg JT, & Fair DA (2019). The heterogeneity problem: Approaches to identify psychiatric subtypes. Trends in cognitive sciences, 23(7), 584–601. 10.1016/j.tics.2019.03.009 [PubMed: 31153774]

Feldman M, Friedler SA, Moeller J, Scheidegger C, & Venkatasubramanian S (2015). Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259–268). Association for Computing Machinery. 10.1145/2783258.2783311

Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, & Chandran D (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Scientific Reports, 8(1), 7426. 10.1038/s41598-018-25773-2 [PubMed: 29743531]

Feuerriegel S, Dolata M, & Schwabe G (2020). Fair AI. Business & Information Systems Engineering, 62(4), 379–384. 10.1007/s12599-020-00650-3

Fiske A, Henningsen P, & Buyx A (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research, 21(5), e13216. 10.2196/13216 [PubMed: 31094356]

Fleming T, Sutcliffe K, Lucassen M, Pine R, & Donkin L (2020). Serious games and gamification in clinical psychology. 10.1016/B978-0-12-818697-8.00011-X

Fletcher RR, Nakeshimana A, & Olubeko O (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health [Methods]. Frontiers in Artificial Intelligence, 3. 10.3389/frai.2020.561802

Fulmer R, Davis T, Costello C, & Joerin A (2021). The ethics of psychological artificial intelligence: Clinical considerations. Counseling and Values, 66(2), 131–144. 10.1002/cvj.12153

Galán CA, Bekele B, Boness C, Bowdring M, Call C, Hails K, McPhee J, Mendes SH, Moses J, Northrup J, Rupert P, Savell S, Sequeira S, Tervo-Clemmens B, Tung I, Vanwoerden S, Womack S, & Yilmaz B (2021). Editorial: A call to action for an antiracist clinical science. Journal of Clinical Child & Adolescent Psychology, 50(1), 12–57. 10.1080/15374416.2020.1860066 [PubMed: 33635185]

Gara MA, Minsky S, Silverstein SM, Miskimen T, & Strakowski SM (2019). A naturalistic study of racial disparities in diagnoses at an outpatient behavioral health clinic. Psychiatric Services, 70(2), 130–134. 10.1176/appi.ps.201800223 [PubMed: 30526340]

Garb HN (2021). Race bias and gender bias in the diagnosis of psychological disorders. Clinical Psychology Review, 90, 102087. 10.1016/j.cpr.2021.102087 [PubMed: 34655834]

Gee D, DeYoung KA, McLaughlin KA, Tillman RM, Barch D, Forbes EE, Krueger R, Strauman TJ, Weierich MR, & Shackman AJ (2021). Training the next generation of clinical psychological scientists: A data-driven call to action. Annual Review of Clinical Psychology. 10.31234/osf.io/xq538

German TP, & Defeyter MA (2000). Immunity to functional fixedness in young children. Psychonomic Bulletin & Review, 7(4), 707–712. 10.3758/BF03213010 [PubMed: 11206213]

Gianfrancesco MA, Tamang S, Yazdany J, & Schmajuk G (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547. 10.1001/jamainternmed.2018.3763 [PubMed: 30128552]

Gorrostieta C, Lotfian R, Taylor K, Brutti R, & Kane J (2019). Gender de-biasing in speech emotion recognition. In Interspeech (pp. 2823–2827). 10.21437/Interspeech.2019-1708

Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H-C, & Jeste DV (2019). Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. Current Psychiatry Reports, 21(11), 116. 10.1007/s11920-019-1094-0 [PubMed: 31701320]

Greenwald AG, & Krieger LH (2006). Implicit bias: Scientific foundations. California Law Review, 94(4), 945–967. 10.2307/20439056

Gruber J, Prinstein MJ, Clark LA, Rottenberg J, Abramowitz JS, Albano AM, Aldao A, Borelli JL, Chung T, Davila J, Forbes EE, Gee DG, Hall GCN, Hallion LS, Hinshaw SP, Hofmann SG, Hollon SD, Joormann J, Kazdin AE, Klein DN, La Greca AM, Levenson RW, MacDonald AW, McKay D, McLaughlin KA, Mendle J, Miller AB, Neblett EW, Nock M, Olatunji BO, Persons JB, Rozek DC, Schleider JL, Slavich GM, Teachman BA, Vine V, & Weinstock LM (2021). Mental health and clinical psychological science in the time of COVID-19: Challenges, opportunities, and a call to action. The American Psychologist, 76(3), 409–426. 10.1037/amp0000707 [PubMed: 32772538]

Hall GCN (2006). Diversity in clinical psychology. Clinical Psychology: Science and Practice, 13(3), 258–262. 10.1111/j.1468-2850.2006.00034.x

Hardeman W, Houghton J, Lane K, Jones A, & Naughton F (2019). A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. International Journal of Behavioral Nutrition and Physical Activity, 16(1), 31. 10.1186/s12966-019-0792-7 [PubMed: 30943983]

Hardt M, Price E, & Srebro N (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems 29 (NIPS 2016).

Hareli S, Kafetsios K, & Hess U (2015). A cross-cultural study on emotion expression and the learning of social norms. Frontiers in Psychology, 6. 10.3389/fpsyg.2015.01501

Harley A (2017). Functional fixedness stops you from having innovative ideas. Nielsen Norman Group. Retrieved February 13 from https://www.nngroup.com/articles/functional-fixedness/

Haselton MG, & Buss DM (2000). Error management theory: A new perspective on biases in cross-sex mind reading. Journal of Personality and Social Psychology, 78(1), 81–91. 10.1037/0022-3514.78.1.81 [PubMed: 10653507]

Hearst MA, Dumais ST, Osuna E, Platt J, & Scholkopf B (1998). Support vector machines. IEEE Intelligent Systems and their Applications, 13(4), 18–28. 10.1109/5254.708428

Hefner A (2002). Knowledge's blind spots: A systems theory perspective on knowledge creation and learning. unpublished work.

Henrich J, Heine SJ, & Norenzayan A (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83. 10.1017/S0140525X0999152X [PubMed: 20550733]

Hindley G, Smeland OB, Frei O, & Andreassen OA (2022). 3 - Big data and the goal of personalized health interventions. In Stein DJ, Fineberg NA, & Chamberlain SR (Eds.), Mental Health in a Digital World (pp. 41–61). Academic Press. 10.1016/B978-0-12-822201-0.00021-6

Hitlin S, & Pinkston K (2013). Values, attitudes, and ideologies: Explicit and implicit constructs shaping perception and action. In Handbook of Social Psychology (pp. 319–339). Springer.

Hochman AL, & Suyemoto KL (2020). Evaluating and dismantling an intervention aimed at increasing White people's knowledge and understanding of racial justice issues. American Journal of Orthopsychiatry, 90(6), 733–750. 10.1037/ort0000506 [PubMed: 32718158]

Hofmann SG, Asnaani A, Vonk IJJ, Sawyer AT, & Fang A (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognitive Therapy and Research, 36(5), 427–440. 10.1007/s10608-012-9476-1 [PubMed: 23459093]

Hooper MW, Nápoles AM, & Pérez-Stable EJ (2020). COVID-19 and racial/ethnic disparities. JAMA, 323(24), 2466–2467. 10.1001/jama.2020.8598 [PubMed: 32391864]

Hoover DS, Vidrine JI, Shete S, Spears CA, Cano MA, Correa-Fernández V, Wetter DW, & McNeill LH (2015). Health literacy, smoking, and health indicators in African American adults. Journal of Health Communication, 20(sup2), 24–33. 10.1080/10810730.2015.1066465 [PubMed: 26513028]

Huff C (2021). Psychology's diversity problem. Monitor on Psychology, 52(7). https://www.apa.org/monitor/2021/10/feature-diversity-problem

IBM Watson Advertising. (2021). How AI is changing advertising. IBM. Retrieved February 13 from https://www.ibm.com/watson-advertising/thought-leadership/how-ai-is-changing-advertising

Iribarren SJ, Cato K, Falzon L, & Stone PW (2017). What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. PLoS ONE, 12(2), e0170581. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5289471/pdf/pone.0170581.pdf [PubMed: 28152012]

Irizarry RA (2020). The role of academia in data science education. Harvard Data Science Review, 2(1). 10.1162/99608f92.dd363929

Jacobson NC, & Bhattacharya S (2022). Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behaviour Research and Therapy, 149, 104013. 10.1016/j.brat.2021.104013 [PubMed: 35030442]

Johnson DDP, Blumstein DT, Fowler JH, & Haselton MG (2013). The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. Trends in Ecology & Evolution, 28(8), 474–481. 10.1016/j.tree.2013.05.014 [PubMed: 23787087]

Johnson KB, Wei W-Q, Weeraratne D, Frisse ME, Misulis K, Rhee K, Zhao J, & Snowdon JL (2021). Precision Medicine, AI, and the Future of Personalized Health Care. Clinical and Translational Science, 14(1), 86–93. 10.1111/cts.12884 [PubMed: 32961010]

Joiner IA (2018). Chapter 1 - Artificial Intelligence: AI is Nearby. In Joiner IA (Ed.), Emerging Library Technologies (pp. 1–22). Chandos Publishing. 10.1016/B978-0-08-102253-5.00002-2

Jones D (2010). A WEIRD View of Human Nature Skews Psychologists' Studies. Science, 328(5986), 1627–1627. 10.1126/science.328.5986.1627 [PubMed: 20576866]

Jones GP, Hickey JM, Di Stefano PG, Dhanjal C, Stoddart LC, & Vasileiou V (2020). Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. arXiv preprint arXiv:2010.03986. 10.48550/arXiv.2010.03986

Jost JT, Rudman LA, Blair IV, Carney DR, Dasgupta N, Glaser J, & Hardin CD (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. Research in Organizational Behavior, 29, 39–69. 10.1016/j.riob.2009.10.001

Juarascio AS, Parker MN, Lagacey MA, & Godfrey KM (2018). Just-in-time adaptive interventions: A novel approach for enhancing skill utilization and acquisition in cognitive behavioral therapy for eating disorders. International Journal of Eating Disorders, 51(8), 826–830. 10.1002/eat.22924 [PubMed: 30051495]

Kamiran F, & Calders T (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. 10.1007/s10115-011-0463-8

Kamiran F, Karim A, & Zhang X (2012). Decision theory for discrimination-aware classification. In 2012 IEEE 12th International Conference on Data Mining (pp. 924–929). 10.1109/ICDM.2012.45

Kamishima T, Akaho S, Asoh H, & Sakuma J (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In ECML PKDD 2012 (pp. 35–50). Springer Berlin Heidelberg. 10.1007/978-3-642-33486-3_3

Kamishima T, Akaho S, & Sakuma J (2011). Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops (pp. 643–650). 10.1109/ICDMW.2011.83

Kamulegeya LH, Okello M, Bwanika JM, Musinguzi D, Lubega W, Rusoke D, Nassiwa F, & Börve A (2019). Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. BioRxiv, 826057. 10.1101/826057

Kehl DL, & Kessler SA (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041

Kehrenberg T, Chen Z, & Quadrianto N (2020). Tuning fairness by balancing target labels [Original Research]. Frontiers in Artificial Intelligence, 3. 10.3389/frai.2020.00033

Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, & Goel S (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684. 10.1073/pnas.1915768117

Krishnapuram B, Williams D, Xue Y, Carin L, Figueiredo MA, & Hartemink AJ (2005). Active learning of features and labels. In Workshop on learning with multiple views at the 22nd International Conference on Machine Learning (ICML-05) (pp. 43–50).

Kuhn M (2009). The caret package. Journal of Statistical Software, 28(5).

Kupers TA (1988). Ending therapy: The meaning of termination. New York University Press.

Kusters R, Misevic D, Berry H, Cully A, Le Cunff Y, Dandoy L, Díaz-Rodríguez N, Ficher M, Grizou J, Othmani A, Palpanas T, Komorowski M, Loiseau P, Moulin Frier C, Nanini S, Quercia D, Sebag M, Soulié Fogelman F, Taleb S, Tupikina L, Sahu V, Vie J-J, & Wehbi F (2020). Interdisciplinary research in artificial intelligence: Challenges and opportunities. Frontiers in Big Data, 3. 10.3389/fdata.2020.577974

Lancet Global Health. (2020). Mental health matters. Lancet Global Health, 8(11), e1352. 10.1016/s2214-109x(20)30432-0 [PubMed: 33069297]

Lancet Global Health Covid 19 Mental Disorders Collaborators. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet Global Health, 398(10312), 1700–1712. 10.1016/s0140-6736(21)02143-7

Lantz B (2019). Machine learning with R: Expert techniques for predictive modeling. Packt Publishing LTD.

Lawrie SM, Fletcher-Watson S, Whalley HC, & McIntosh AM (2019). Predicting major mental illness: ethical and practical considerations. BJPsych open, 5(2), e30–e30. 10.1192/bjo.2019.11 [PubMed: 31068241]

Le Glaz A, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, & Lemey C (2021). Machine learning and natural language processing in mental health: Systematic review. Journal of Medical Internet Research, 23(5), e15708. 10.2196/15708 [PubMed: 33944788]

Liao P, Dempsey W, Sarker H, Hossain SM, al'Absi M, Klasnja P, & Murphy S (2018). Just-in-Time but Not Too Much: Determining Treatment Timing in Mobile Health. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol, 2(4), Article 179. 10.1145/3287057 [PubMed: 30801052]

Ling CX, & Sheng VS (2010). Class Imbalance Problem. In Sammut C & Webb GI (Eds.), Encyclopedia of Machine Learning (pp. 171–171). Springer US. 10.1007/978-0-387-30164-8_110

Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, & Puri R (2019). Bias mitigation post-processing for individual and group fairness. ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2847–2851. 10.48550/arXiv.1812.06135

Lombardo MV, Lai M-C, & Baron-Cohen S (2019). Big data approaches to decomposing heterogeneity across the autism spectrum. Molecular Psychiatry, 24(10), 1435–1450. 10.1038/s41380-018-0321-0 [PubMed: 30617272]

Lum K, & Isaac W (2016). To predict and serve? Significance, 13(5), 14–19. 10.1111/j.1740-9713.2016.00960.x

Luong BT, Ruggieri S, & Turini F (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA. 10.1145/2020408.2020488

Luxton DD (2014). Artificial intelligence in psychological practice: Current and future applications and implications. Professional Psychology: Research and Practice, 45(5), 332.

Maguire A (2016). Interaction Institute for Social Change | Artist: Angus Maguire. interactioninstitute.org and madewithangus.com

Marlatt GA, Bowen SW, & Witkiewitz K (2009). Chapter eleven - Relapse prevention: evidence base and future directions. In Miller PM (Ed.), Evidence-Based Addiction Treatment (pp. 215–232). Academic Press. 10.1016/B978-0-12-374348-0.00011-2

Martin LR, Williams SL, Haskard KB, & Dimatteo MR (2005). The challenge of patient adherence. Therapeutics and Clinical Risk Management, 1(3), 189–199. https://pubmed.ncbi.nlm.nih.gov/18360559 [PubMed: 18360559]

Maura J, & de Mamani AW (2017). Mental health disparities, treatment engagement, and attrition among racial/ethnic minorities with severe mental illness: A review. Journal of Clinical Psychology in Medical Settings, 24(3–4), 187–210. 10.1007/s10880-017-9510-2 [PubMed: 28900779]

McGuire TG, & Miranda J (2008). New evidence regarding racial and ethnic disparities in mental health: policy implications. Health Affairs (Project Hope), 27(2), 393–403. 10.1377/hlthaff.27.2.393 [PubMed: 18332495]

McKnight-Eily LR, Okoro CA, Strine TW, Verlenden J, Hollis ND, Njai R, Mitchell EW, Board A, Puddy R, & Thomas C (2021). Racial and ethnic disparities in the prevalence of stress and worry, mental health conditions, and increased substance use among adults during the COVID-19 pandemic—United States, April and May 2020. Morbidity and Mortality Weekly Report, 70(5), 162. 10.15585/mmwr.mm7005a [PubMed: 33539336]

Mehrabi N, Morstatter F, Saxena N, Lerman K, & Galstyan A (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6). 10.1145/3457607

Miconi D, Li ZY, Frounfelker RL, Santavicca T, Cénat JM, Venkatesh V, & Rousseau C (2021). Ethno-cultural disparities in mental health during the COVID-19 pandemic: a cross-sectional study on the impact of exposure to the virus and COVID-19-related discrimination and stigma on mental health across ethno-cultural groups in Quebec (Canada). BJPsych open, 7(1). 10.1192/bjo.2020.146

Miranda J, McGuire TG, Williams DR, & Wang P (2008). Mental health in the context of health disparities. American Journal of Psychiatry, 165(9), 1102–1108. 10.1176/appi.ajp.2008.08030333 [PubMed: 18765491]

Mulligan DK, Kroll JA, Kohli N, & Wong RY (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–36. 10.1145/3359221 [PubMed: 34322658]

Murphy J, & Hallinger P (1989). Equity as access to learning: curricular and instructional treatment differences. Journal of Curriculum Studies, 21(2), 129–149. 10.1080/0022027890210203

Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, & Murphy SA (2018). Just-in-time adaptive interventions (JITAIS) in mobile health: Key components and design principles for ongoing health behavior support. Annals of Behavioral Medicine, 52(6), 446–462. 10.1007/s12160-016-9830-8 [PubMed: 27663578]

National Academies of Sciences, E., and Medicine. (2016). The promises and perils of digital strategies in achieving health equity: Workshop summary. National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK373441/

National Institute of Mental Health. (2017). Technology and the future of mental health treatment. https://www.nimh.nih.gov/health/topics/technology-and-the-future-of-mental-health-treatment

National Institute of Mental Health. (2022). Mental Illness. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/mental-illness

Newman CF (2012). Maintaining treatment gains and planning for termination. Core Competencies in Cognitive-Behavioral Therapy: Becoming a Highly Effective and Competent Cognitive-Behavioral Therapist, 181.

Ngiam KY, & Khor IW (2019). Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262–e273. 10.1016/S1470-2045(19)30149-4 [PubMed: 31044724]

Noah B, Keller MS, Mosadeghi S, Stein L, Johl S, Delshad S, Tashjian VC, Lew D, Kwan JT, & Jusufagic A (2018). Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. NPJ Digital Medicine, 1(1), 1–12. [PubMed: 31304287]

Noor P (2020). Can we trust AI not to further embed racial bias and prejudice? BMJ, 368. 10.1136/bmj.m363

Obermeyer Z, Powers B, Vogeli C, & Mullainathan S (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. 10.1126/science.aax2342 [PubMed: 31649194]

Office of the Surgeon General. (2001). Culture counts: The influence of culture and society on mental health. Mental Health: Culture, Race, and Ethnicity: A Supplement to Mental Health: A Report of the Surgeon General. Substance Abuse and Mental Health Services Administration (US). https://www.ncbi.nlm.nih.gov/books/NBK44249/

Oliver A, & Mossialos E (2004). Equity of access to health care: Outlining the foundations for action. Journal of Epidemiology and Community Health, 58(8), 655. 10.1136/jech.2003.017731 [PubMed: 15252067]

Ornell F, Borelli WV, Benzano D, Schuch JB, Moura HF, Sordi AO, Kessler FHP, Scherer JN, & von Diemen L (2021). The next pandemic: impact of COVID-19 in mental healthcare assistance in a nationwide epidemiological study. Lancet Reg Health Am, 4, 100061. 10.1016/j.lana.2021.100061 [PubMed: 34518824]

Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF 3rd, & Aizenstein HJ (2015). Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. International Journal of Geriatric Psychiatry, 30(10), 1056–1067. 10.1002/gps.4262 [PubMed: 25689482]

Payne BK, & Hannay JW (2021). Implicit bias reflects systemic racism. Trends in Cognitive Sciences. 10.1016/j.tics.2021.08.001

Payne BK, Vuletich HA, & Lundberg KB (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. Psychological Inquiry, 28(4), 233–248. 10.1080/1047840X.2017.1335568

Perry BL, Aronson B, & Pescosolido BA (2021). Pandemic precarity: COVID-19 is exposing and exacerbating inequalities in the American heartland. Proceedings of the National Academy of Sciences, 118(8), e2020685118. 10.1073/pnas.2020685118

Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, & Businelle MS (2021). Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction. 10.1111/add.15687

Petersen F, Mukherjee D, Sun Y, & Yurochkin M (2021). Post-processing for individual fairness. Advances in Neural Information Processing Systems, 34.

Piaget J (1952). The fourth stage: The coordination of the secondary schemata and their application to new situations. In The Origins of Intelligence in Children (pp. 210–262, 419 Pages). W W Norton & Co, New York, NY. 10.1037/11494-005

Pleiss G, Raghavan M, Wu F, Kleinberg J, & Weinberger KQ (2017). On fairness and calibration. Advances in Neural Information Processing Systems, 30. https://doi.org/arXiv:1709.02012

Pramana G, Parmanto B, Lomas J, Lindhiem O, Kendall PC, & Silk J (2018). Using mobile health gamification to facilitate cognitive behavioral therapy skills practice in child anxiety treatment: Open clinical trial. JMIR Serious Games, 6(2), e9. 10.2196/games.8902 [PubMed: 29748165]

Price M, Yuen EK, Goetter EM, Herbert JD, Forman EM, Acierno R, & Ruggiero KJ (2014). mHealth: a mechanism to deliver more accessible, more effective mental health care. Clinical Psychology & Psychotherapy, 21(5), 427–436. 10.1002/cpp.1855 [PubMed: 23918764]

Purtle J (2020). COVID-19 and mental health equity in the United States. Social psychiatry and psychiatric epidemiology, 55(8), 969–971. 10.1007/s00127-020-01896-8 [PubMed: 32556376]

Quadrianto N, & Sharmanska V (2017). Recycling privileged learning and distribution matching for fairness. Advances in Neural Information Processing Systems, 30, 677–688.

Queirós A, Alvarelhão J, Cerqueira M, Silva AG, Santos M, & Pacheco Rocha N (2018). Remote care technology: a systematic review of reviews and meta-analyses. Technologies, 6(1), 22. 10.3390/technologies6010022

Quinlan JR (1990). Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20(2), 339–346. 10.1109/21.52545

Radfar A, Ferreira MM, Sosa JP, & Filip I (2021). Emergent crisis of Covid-19 Pandemic: Mental health challenges and opportunities [Review]. Frontiers in Psychiatry, 12. 10.3389/fpsyt.2021.631008

Ramos G, Ponting C, Labao JP, & Sobowale K (2021). Considerations of diversity, equity, and inclusion in mental health apps: A scoping review of evaluation frameworks. Behaviour Research and Therapy, 147, 103990. 10.1016/j.brat.2021.103990 [PubMed: 34715396]

Raue PJ, & Sirey JA (2011). Designing personalized treatment engagement interventions for depressed older adults. The Psychiatric clinics of North America, 34(2), 489–x. 10.1016/j.psc.2011.02.011 [PubMed: 21536170]

Richardson R, Schultz JM, & Crawford K (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. NYU Law Review, 94, 15. https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/

Roberts SO, & Rizzo MT (2021). The psychology of American racism. American Psychologist, 76(3), 475–487. 10.1037/amp0000642 [PubMed: 32584061]

Rodriguez JD, Perez A, & Lozano JA (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3), 569–575. 10.1109/TPAMI.2009.187 [PubMed: 20075479]

Rozental A, Andersson G, & Carlbring P (2019). In the absence of effects: An individual patient data meta-analysis of non-response and its predictors in internet-based cognitive behavior therapy [Systematic Review]. Frontiers in Psychology, 10. 10.3389/fpsyg.2019.00589

Sali AW, Anderson BA, & Courtney SM (2018). Information processing biases in the brain: Implications for decision-making and self-governance. Neuroethics, 11(3), 259–271. 10.1007/s12152-016-9251-1 [PubMed: 30555600]

Saltzman LY, Lesen AE, Henry V, Hansel TC, & Bordnick PS (2021). COVID-19 mental health disparities. Health security, 19(S1), S-5–S-13. 10.1089/hs.2021.0017 [PubMed: 34014118]

Sap M, Card D, Gabriel S, Choi Y, & Smith AN (2019). The risk of racial bias in hate speech detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistic,

Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, & Ghassemi M (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176–2182. 10.1038/s41591-021-01595-0

Shaheen MY (2021). AI in Healthcare: medical and socio-economic benefits and challenges. ScienceOpen Preprints.

Shatte ABR, Hutchinson DM, & Teague SJ (2019). Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine, 49(9), 1426–1448. 10.1017/S0033291719000151 [PubMed: 30744717]

Sica GT (2006). Bias in research studies. Radiology, 238(3), 780–789. 10.1148/radiol.2383041109 [PubMed: 16505391]

Sneed RS, Key K, Bailey S, & Johnson-Lawrence V (2020). Social and psychological consequences of the COVID-19 pandemic in African-American communities: Lessons from Michigan. *Psychological Trauma: Theory, Research*, Practice, and Policy, 12(5), 446–448. 10.1037/tra0000881

Stewart DW, Cano MÁ, Correa-Fernández V, Spears CA, Li Y, Waters AJ, Wetter DW, & Vidrine JI (2014). Lower health literacy predicts smoking relapse among racially/ethnically diverse smokers with low socioeconomic status. BMC Public Health, 14(1), 716. 10.1186/1471-2458-14-716 [PubMed: 25018151]

Straw I, & Callison-Burch C (2020). Artificial Intelligence in mental health and the biases of language based models. PLoS ONE, 15(12), e0240376. 10.1371/journal.pone.0240376 [PubMed: 33332380]

Sturm S (2006). The architecture of inclusion: Advancing workplace equity in higher education. Harv. JL & Gender, 29, 247.

Sussman LK, Robins LN, & Earls F (1987). Treatment-seeking for depression by black and white Americans. Social science & medicine, 24(3), 187–196. 10.1016/0277-9536(87)90046-3 [PubMed: 3824001]

Sutton CD (2005). Classification and regression trees, bagging, and boosting. Handbook of statistics, 24, 303–329.

Techjury.net. (2019). Infographic: How AI is being deployed across industries. Robotics Business Review. Retrieved February 13 from https://www.roboticsbusinessreview.com/ai/infographic-how-ai-is-being-deployed-across-industries/

Thapa S, & Nielsen JB (2021). Association between health literacy, general psychological factors, and adherence to medical treatment among Danes aged 50–80 years. BMC Geriatrics, 21(1), 386. 10.1186/s12877-021-02339-y [PubMed: 34174815]

Tian C-J, Lv J, & Xu X-F (2021). Evaluation of feature selection methods for mammographic breast cancer diagnosis in a unified framework. BioMed Research International, 2021, 6079163–6079163. 10.1155/2021/6079163 [PubMed: 34646886]

Topol EJ (2020). Welcoming new guidelines for AI clinical research. Nature Medicine, 26(9), 1318–1320. 10.1038/s41591-020-1042-x

Torous J, Jän Myrick K, Rauseo-Ricupero N, & Firth J (2020). Digital Mental Health and COVID-19: Using Technology Today to Accelerate the Curve on Access and Quality Tomorrow. JMIR Mental Health, 7(3), e18848. 10.2196/18848 [PubMed: 32213476]

Trotter M (2021). What is covariate shift. Seldon. https://www.seldon.io/what-is-covariate-shift/

Turpin G, & Coleman G (2010). Clinical psychology and diversity: Progress and continuing challenges. Psychology Learning & Teaching, 9(2), 17–27. 10.2304/plat.2010.9.2.17

Varshney U (2007). Pervasive healthcare and wireless health monitoring. Mobile Networks and Applications, 12(2–3), 113–127.

Veale M, Kleek MV, & Binns R (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. Paper 440). Association for Computing Machinery. 10.1145/3173574.3174014

Vogels EA (2021). Digital divide persists even as Americans with lower incomes make gains in tech adoption. Pew Research Center. Retrieved Febuary 3, 2022 from https://pewrsr.ch/2TRM7cP

Wahle F, Kowatsch T, Fleisch E, Rufer M, & Weidt S (2016). Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. JMIR mhealth Uhealth, 4(3), e111. 10.2196/mhealth.5960 [PubMed: 27655245]

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, & Walker-Smith JA (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. The Lancet, 351(9103), 637–641. 10.1016/S0140-6736(97)11096-0

Wang L, & Miller LC (2020). Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. Health Communication, 35(12), 1531–1544. 10.1080/10410236.2019.1652388 [PubMed: 31488002]

Warner B, & Misra M (1996). Understanding Neural Networks as Statistical Tools. The American Statistician, 50(4), 284–293. 10.1080/00031305.1996.10473554

Warner R (1979). Racial and sexual bias in psychiatric diagnosis: Psychiatrists and other mental health professionals compared by race, sex, and discipline. Journal of Nervous and Mental Disease, 167(5), 303–310. 10.1097/00005053-197905000-00007 [PubMed: 448333]

Warren R, Carlisle K, Mihala G, & Scuffham PA (2018). Effects of telemonitoring on glycaemic control and healthcare costs in type 2 diabetes: A randomised controlled trial. Journal of Telemedicine and Telecare, 24(9), 586–595. https://journals.sagepub.com/doi/10.1177/1357633X17723943?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed [PubMed: 28814128]

Weisz JR, Weiss B, Han SS, Granger DA, & Morton T (1995). Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies. Psychological Bulletin, 117(3), 450–468. 10.1037/0033-2909.117.3.450 [PubMed: 7777649]

Widiger TA, & Spitzer RL (1991). Sex bias in the diagnosis of personality disorders: Conceptual and methodological issues. Clinical Psychology Review, 11(1), 1–22. 10.1016/0272-7358(91)90135-H

Williams MT (2019). Adverse racial climates in academia: Conceptualization, interventions, and call to action. New Ideas in Psychology, 55, 58–67. 10.1016/j.newideapsych.2019.05.002

Woodworth B, Gunasekar S, Ohannessian MI, & Srebro N (2017). Learning non-discriminatory predictors. In Satyen K & Ohad S (Eds.), Proceedings of the 2017 Conference on Learning Theory (Vol. 65, pp. 1920–1953). PMLR. https://proceedings.mlr.press/v65/woodworth17a.html

World Health Organization. (2020). COVID-19 disrupting mental health services in most countries, WHO survey. https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey

Yamamoto N, Ochiai K, Inagaki A, Fukazawa Y, Kimoto M, Kiriu K, Kaminishi K, Ota J, Okimura T, Terasawa Y, & Maeda T (2018). Physiological stress level estimation based on smartphone logs. In 2018 Eleventh International Conference on Mobile Computing and Ubiquitous Network (ICMU) (pp. 1–6). 10.23919/ICMU.2018.8653590

Yarber AL (2022). What is Bias? ADVANCE Geo Partnership. https://serc.carleton.edu/advancegeo/resources/bias.html

Yeom S, & Fredrikson M (2020). Individual fairness revisited: transferring techniques from adversarial robustness. ArXiv, abs/2002.07738. 10.48550/arXiv.2002.07738

Zafar MB, Valera I, Rodriguez MG, & Gummadi KP (2017a). Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web (pp. 1171–1180). International World Wide Web Conferences Steering Committee. 10.1145/3038912.3052660

Zafar MB, Valera I, Rodriguez MG, & Gummadi KP (2017b). Fairness constraints: Mechanisms for fair classification. In Aarti S & Jerry Z (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Vol. 54, pp. 962–970). PMLR. https://proceedings.mlr.press/v54/zafar17a.html

Zhang AY, Snowden LR, & Sue S (1998). Differences between Asian and White Americans' help seeking and utilization patterns in the Los Angeles area. Journal of Community Psychology, 26(4), 317–326. 10.1002/(SICI)1520-6629(199807)26:4&lt;317::AID-JCOP2&gt;3.0.CO;2-Q

Zhang BH, Lemoine B, & Mitchell M (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340). Association for Computing Machinery. 10.1145/3278721.3278779

Zhao J, Zhou Y, Zeyu L, Wang W, Chang K (2018). Learning gender-neutral word embeddings. arXiv preprint ArXiv: 1809.01496. 10.48550/arXiv.1809.01496

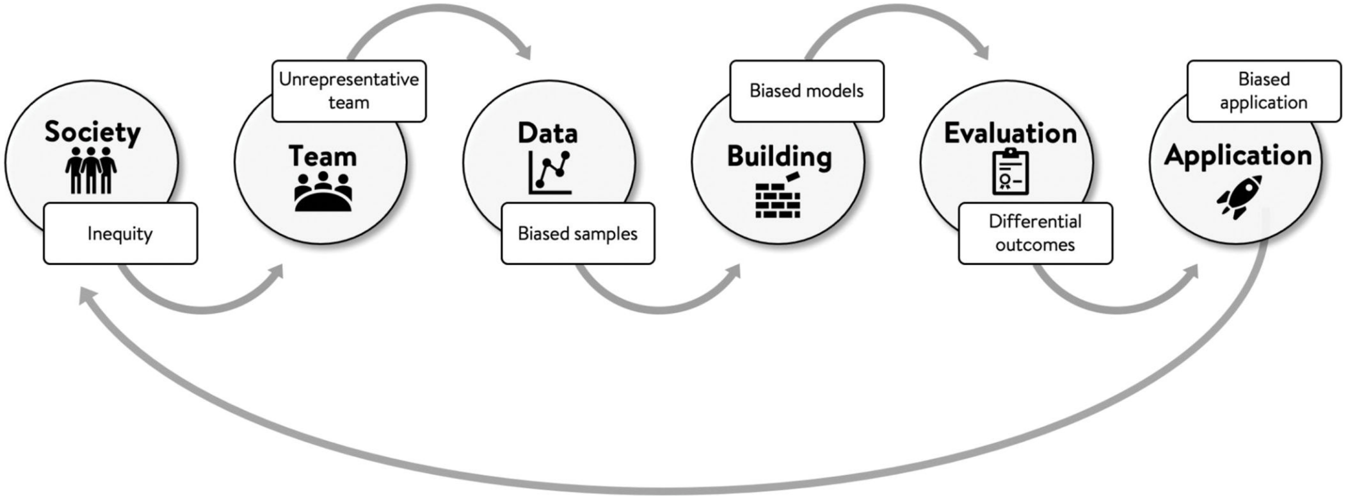

Figure 1.

Overview of the model-building process and points at which bias can be introduced to models.

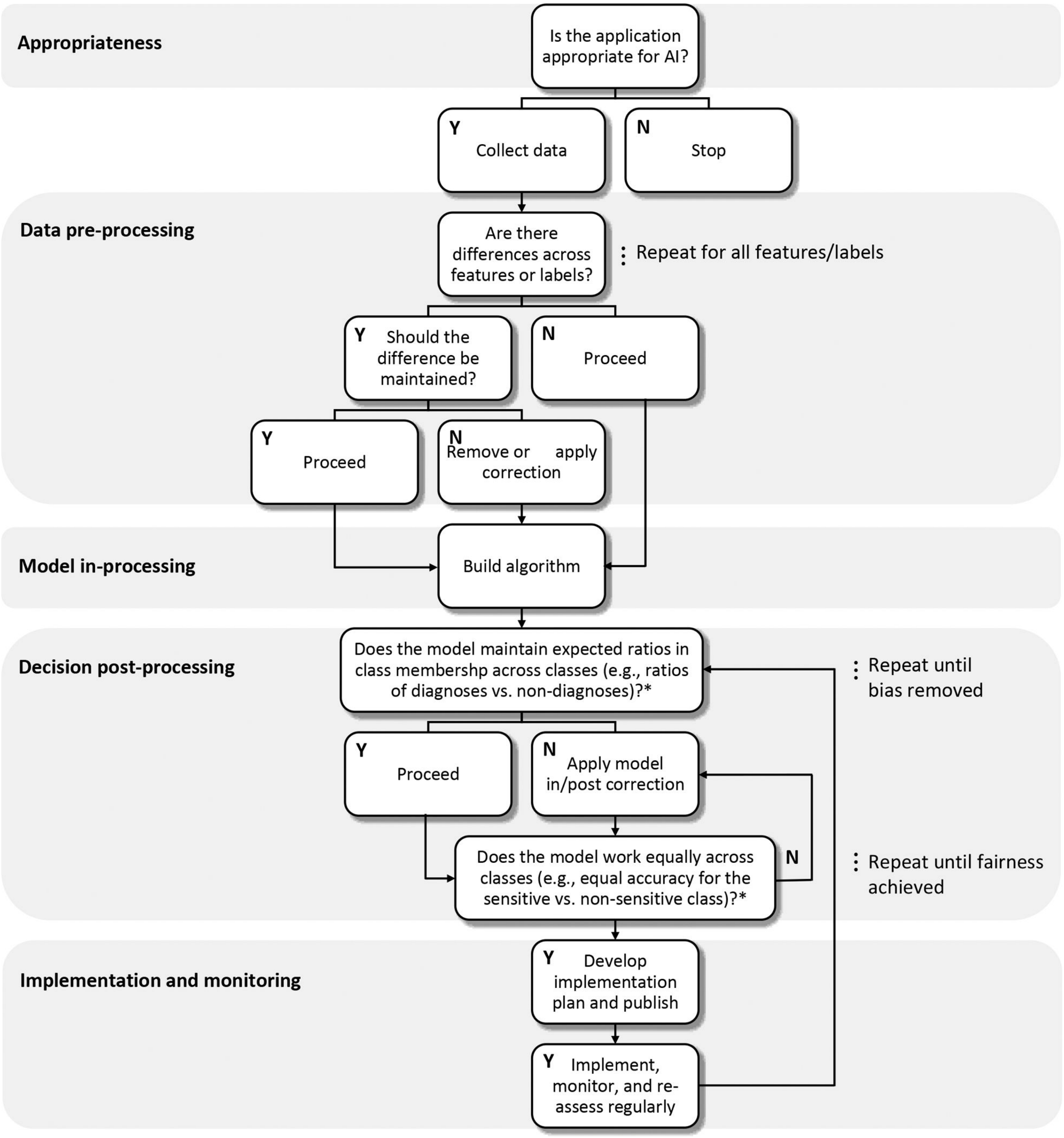
Appropriateness
Is the application appropriate for AI?
Y Collect data
N Stop
Data pre-processing
Are there differences across features or labels?
Repeat for all features/labels
Y Should the difference be maintained?
N Proceed
Y Proceed
N Remove or apply correction
Model in-processing
Build algorithm
Decision post-processing
Does the model maintain expected ratios in class membershp across classes (e.g., ratios of diagnoses vs. non-diagnoses)?*
Repeat until bias removed
Y Proceed
N Apply model in/post correction
Does the model work equally across classes (e.g., equal accuracy for the sensitive vs. non-sensitive class)?*
N
Repeat until fairness achieved
Implementation and monitoring
Y Develop implementation plan and publish
Y Implement, monitor, and re-assess regularly

Figure 2.
Overview of the steps for incorporating bias in AI assessment and mitigation. Y = *yes*. N = *no*. Stakeholders/target population should be interviewed at each stage and feedback incorporated. *The questions presented in the decision post-processing stage can be completed simultaneously or the order can be interchanged.

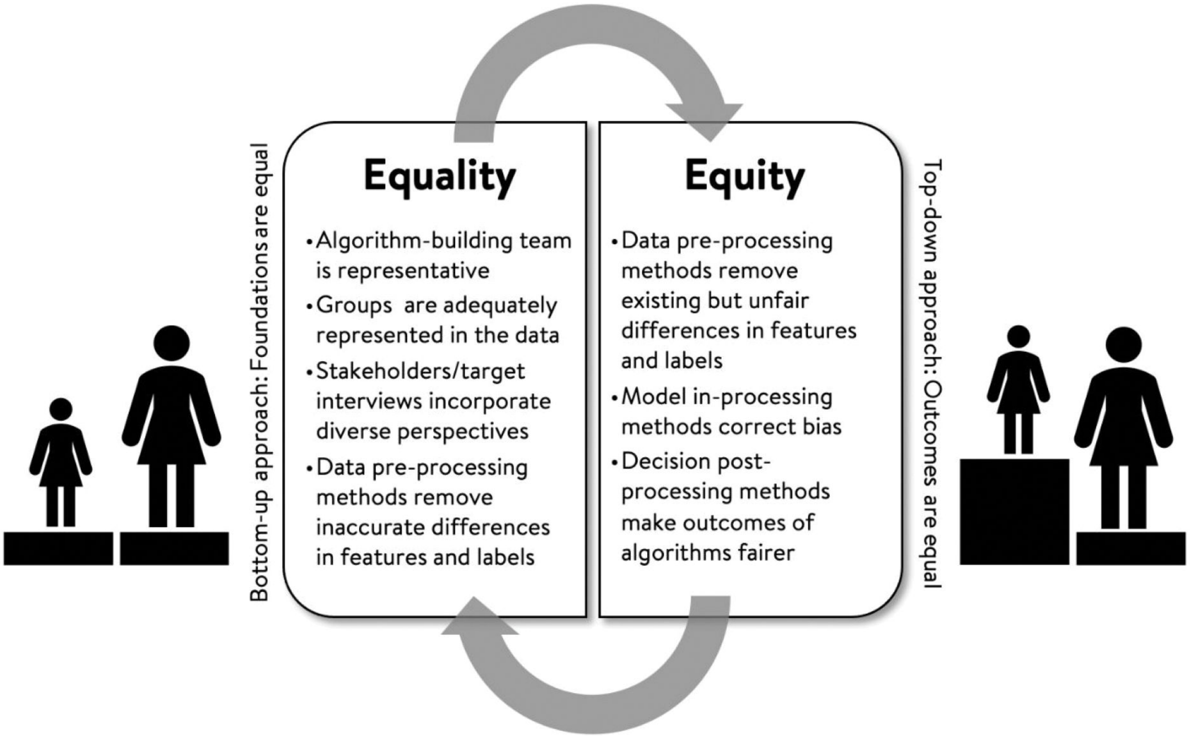
Equality
• Algorithm-building team is representative
• Groups are adequately represented in the data
• Stakeholders/target interviews incorporate diverse perspectives
• Data pre-processing methods remove inaccurate differences in features and labels
Equity
• Data pre-processing methods remove existing but unfair differences in features and labels
• Model in-processing methods correct bias
• Decision post-processing methods make outcomes of algorithms fairer
Bottom-up approach: Foundations are equal
Top-down approach: Outcomes are equal

Figure 3.

Visual depiction demonstrating how strategies to assess and mitigate bias in AI integrate with equality- and equity-based approaches to address disparities in mental health. Equality-equity visualization adapted to bias in AI from the Interaction for Social Change Model (Maguire, 2016).

Table 1

Summary of Types of Bias that Can Influence the Development and Application of AI Models

Bias Type	Definition	Example
Sociocultural Foundations		
Structural inequalities	Institutions, governments, and organizations have inherent biases that cause some groups (e.g., race, gender, age) to be favored over others; can be in the form of education, housing, health care, environment, media, etc.	Compared to children in wealthier areas, children living in under-resourced neighborhoods often receive fewer educational resources, consume lower quality air and water, and have poorer municipal services (Evans, 2004). It is an important preliminary step to consider how underlying structural inequalities might impact the model development process because such foundations can unconsciously impact our perceptions and decision-making. For example, children without access to high-quality educational resources may have lower test scores. If we aim to build algorithms to predict important outcomes, then an important foundational step would be to consider how structural inequalities might relate to the outcome and thus lead to biased results (Churchwell et al., 2020)
Historical bias	Bias in data resulting from the context of a given historical time.	Algorithms built from Google News articles exhibited bias by over-associating certain occupations with men versus women (e.g., men as computer programmers and women as homemakers). Because the algorithms were built from historically sexist text data, the bias was carried forward to the resulting algorithm (Bolukbaski et al., 2016).
Homogeneous teams	Teams lacking diversity can increase the chances of implicit biases going unnoticed because the members of the team may all may share the same biases.	Scientific explanations of human fertilization have traditionally considered sperm to be “active” and “competitive” and eggs to be “passive” and “receptive” (Campo-Engelstein & Jonson, 2013). However, new evidence shows that these prior explanations of human fertilization are inaccurate, with eggs playing a more active role in determining which sperm is fertilized than previously thought (Fitzpatrick, 2020). Members of other genders could have helped to identify biased thinking in the conceptualization and theory of human fertilization.
Data Collection		
Representation bias	Occurs when the collected sample is not diverse enough to represent the population for whom an application is made.	In automated speech recognition systems from five of the top tech companies, there is a 35% word-error rate on average for Black and African American speakers compared to 19% for non-Hispanic White speakers, demonstrating that the input dataset was not representative enough to account for the speech of people from different cultural and racial and ethnic backgrounds (Koenecke, 2020; Bajorek, 2019).
Measurement bias	Occurs when the measures and features researchers decide to use when building models are more accurate in some groups than others.	Algorithms might use automated speech transcription as a feature to predict diagnosis. However, the feature itself is biased because it produces more word-detection errors in women and minoritized groups compared to non-Hispanic White men (Bajorek, 2019). Additionally, measurement biases can often arise in age-heterogeneous samples because tests developed for young adults may not work for older adults (Zeidner, 1987).
Aggregation bias	Occurs when a target group is treated as a monolith and false conclusions are made because the model did not account for group diversity.	Hemoglobin AIc is regularly used to diagnose and monitor diabetes. However, research shows that levels of Hemoglobin AIc differ across ethnicities, contributing to bias in diagnostic classification (Ford et al., 2019).
Model Building		
Confirmation bias	Occurs when researchers favor information that supports pre-existing beliefs, causing them to avoid looking for information that may be to the contrary.	Wakefield et al. (1998) inaccurately linked the MMR vaccine to autism. Though the study was retracted from the *British Medical Journal* in 2010 after evidence that Wakefield manipulated and ignored much of the data, the disproven claim still impacts community perception of vaccines today (Belluz, 2019; Deer, 2011).
Label bias	In supervised machine learning, labels must be applied to the training data and then fed to the machine learning algorithm so that the AI can properly predict what future values will be. However, these labels may not always represent all the possible	Individuals were tasked with identifying and labeling Twitter posts as hate speech for a machine learning algorithm. Inaccuracy in labeling the tweets of African American people directly led to the algorithm’s decision to inaccurately label future African Americans’ tweets as hate speech (Davidson et al., 2019; Sap et al., 2019).

Bias Type	Definition	Example
	labels for a given variable or can be inaccurate, resulting in biased predictions.	
Feature selection bias	In classification models, feature selection is a technique used to select the variables that will be the best predictors for a given target value or to reduce the number of unrelated variables that may influence predictions. Bias in this technique occurs when unrepresentative or inaccurate features are selected.	In mammogram interpretation, images are difficult to interpret and must be cleaned to draw accurate conclusions. Removal of irrelevant image information is necessary to denoise the image. Sometimes the images may contain additional content that is insufficient to be used as features. Failure to exclude such noise could result in algorithms producing poor generalizations across different images that may contain a variety of irrelevant background content (Tian et al., 2021).
Model Performance and Evaluation		
Class imbalance bias	Occurs when facet *a* has fewer training samples compared to facet *b* in the dataset; this results in models preferentially fitting larger facets at the expense of smaller facets, which can result in greater error rates for facet a. Models are also higher risk of overfitting smaller datasets, which can cause larger test error for facet *a*.	If a machine learning model is trained primarily on data from middle-aged individuals, it might be less accurate when making predictions involving younger and older people. Similarly, when there are very few examples in the dataset of people from minoritized backgrounds compared to the many examples for the majority, predictions for minorities may be less accurate (Amazon Web Services; Ling & Sheng, 2010).
Covariate shift	When the test data distribution does not match the distribution of the training data due to shifts in the target population with time.	Covariate shift can occur for image categorization and facial recognition where models achieve high accuracy on a labeled training dataset, but model accuracy can decrease when deployed with live data. For example, subtle changes in lighting could shift data distribution points lowering model accuracy (Trotter, 2021). Light quality and intensity are covariates that impact the relationship between the features and labels. If light quality and intensity change over time (e.g., changes in the time of day the picture was taken or the type and quality of lights used in the home or office), the algorithm performance can be impacted and rendered less accurate over time.
Evaluation bias	When a model demonstrates false efficiency because the training set used is unrepresentative of a minority group and then later demonstrates efficiency on the test set due to it being equally unrepresentative. Thus, showing accuracy but false generalizability.	The use of inappropriate and disproportionate benchmarks for evaluation of application in imbalanced datasets that contain mainly lighter-complexion individuals; these benchmarks are used in the evaluation of facial recognition systems that are biased on skin color and gender (Buolamwini & Gebru, 2018; Mehrabi et al., 2021)
Human Inference and Deployment		
Deployment bias	Occurs when a model is deployed in a setting or context for which it was not designed.	Models developed to predict risk of recidivism are often used to determine a defendant's length of stay, which was not the model's intended use (Kehl & Kessler, 2017).
User interaction bias	Occurs when users' behavior and interaction with a given application or website can introduce bias into the algorithm.	YouTube's restricted mode was censoring LGBTQIA+ content. They used user input to inform an automated detection algorithm to determine whether content was appropriate. LGBTQIA+ content, even though it was not explicit, was censored due to people flagging such videos. The algorithm used that information and thus classified similar videos as inappropriate. This issue represents how user interaction and the dominant society's moral values could further marginalize historically marginalized groups (Bensinger & Albergotti, 2019).
Feedback loop bias	Occurs when a model with an existing bias further reinforces that bias via human-application interaction and the use of newly collected data to feed back into the algorithm for further prediction and training.	Predictive policing models utilize arrest data as training datasets. If police are more likely to make more arrests in a more heavily policed area, using arrest data to predict crime hotspots will disproportionally channel policing efforts to already over-policed communities (Chouldechova & Roth, 2018; Lum & Isaac, 2016).

Note. The list of bias types that can affect AI is not exhaustive.

Table 2

Recommendations for Assessing and Mitigating Bias in Mental Health AI Applications

Model Building	
1.	Recruit a diverse team to build the algorithms (e.g., Cogwill et al., 2020).
2.	Recruit stakeholders and representatives from target population to inform all stages of development (e.g., Lee et al., 2019).
3.	Create a standardized system for eliciting feedback and revising models, including standard questions (e.g., Mulligan et al., 2019).
4.	Elicit feedback from stakeholders/target population on problem/solution conceptualization, revise (e.g., Lee et al., 2019; Smith & Rustagi, 2020).
5.	Elicit feedback from stakeholders/target population on features and labels to be used, revise (e.g., Lee et al., 2019; Smith & Rustagi, 2020).
6.	Collect representative data that matches the target population and application to be implemented (e.g., Buloamwini & Gebru, 2018; Hupoint & Fernandez, 2019; Karkkainen & Joo, 2019; Merler et al., 2019).
7.	If possible, avoid using sensitive attributes as features in model development (e.g., Kilbertus et al., 2018; Yan et al., 2020).
8.	After collecting data, conduct a pre-processing bias assessment on features and labels (e.g., Celis et al., 2016; Celis et al., 2018; Zhang et al., 2018).
9.	Elicit feedback from stakeholders/target population on data pre-processing assessment, revise (e.g., Lee et al., 2019; Smiith & Rustagi, 2020).
10.	If possible, choose interpretable and intuitive models.
Model Evaluation	
1.	Examine ratios of predictions (e.g., ratio of diagnoses versus non-diagnoses) across sensitive attributes and combinations thereof (e.g., Hardt et al. 2016).
2.	Examine model performance (e.g., accuracy, kappa) across sensitive attributes and combinations thereof (e.g., Hardt et al. 2016).
3.	Elicit feedback from stakeholders/target population on decision post-processing assessment, revise (e.g., Hardt et al. 2016).
Bias Mitigation	
1.	If bias is detected, apply model in-processing and decision post-processing methods (e.g., Feng, 2022; Kamishima et al., 2012; Mehrabi et al., 2022; Oneto et al00., 2019; Pfohl et al., 2021; Ustun, 2019; Zemel et al., 2013; Zhao et al., 2018).
2.	Repeat Model Evaluation Steps 1–3 until bias is removed from the model.
Model Implementation	
1.	Identify and plan for appropriate use cases for applying the algorithm.
2.	Identify and plan for worst-case scenarios and outline remediation plans for these scenarios (e.g., High-Level Expert Group on Artificial Intelligence, 2020).
3.	Delineate and implement safeguards and model monitoring parameters (e.g., High-Level Expert Group on Artificial Intelligence, 2020).
4.	Delineate and implement opt-out and appeal processes that are easy and straightforward (e.g., Schwartz et al. 2020).
5.	Elicit feedback from stakeholders/target population on model results and implementation plan, revise.
6.	Publish algorithm, de-identified dataset, documentation, and Bias and Fairness Assessment results (e.g., Shin 2021).
7.	Maintain regular monitoring and assessment of algorithm impact and update the model as needed (e.g., Schwartz et al. 2020).
8.	Elicit feedback from stakeholders/target population on monitoring and impact assessments, revise.
9.	Repeat process for any model adaptations to new target populations or use cases.