2. AI Bias: Context and Terminology

For purposes of this publication, the term Artificial Intelligence (AI) refers to a large class of software-based systems that receive signals from the environment and take actions that affect that environment by generating outputs such as content, predictions, recommendations, classifications, or decisions influencing the environments they interact with, among other outputs [63]. Machine learning (ML) refers more specifically to the "field of study that gives computers the ability to learn without being explicitly programmed," [64] or to computer programs that utilize data to learn and apply patterns or discern statistical relationships. Common ML approaches include, but are not limited to, regression, random forests, support vector machines, and artificial neural networks. ML programs may or may not be used to make predictions of future events. ML programs also may be used to create input for additional ML programs. AI includes ML within its scope.

While AI holds great promise, the convenience of automated classification and discovery within large datasets can come with significant downsides to individuals and society through the amplification of existing biases. Bias can be introduced purposefully or inadvertently into an AI system, or it can emerge as the AI is used in an application. Some types of AI bias are purposeful and beneficial. For example, the ML systems that underlie AI applications often model our implicit biases with the intent of creating positive experiences for online shopping or identifying content of interest [65, 66]. The proliferation of recommender systems and other modeling and predictive approaches has also helped to expose the many negative social biases baked into these processes, which can reduce public trust [67–70].

AI is neither built nor deployed in a vacuum, sealed off from societal realities of discrimination or unfair practices. Understanding AI as a socio-technical system acknowledges that the processes used to develop technology are more than their mathematical and computational constructs. A socio-technical approach to AI takes into account the values and behavior modeled from the datasets, the humans who interact with them, and the complex organizational factors that go into their commission, design, development, and ultimate deployment.

2.1 Characterizing AI bias

2.1.1 Contexts for addressing AI bias

Statistical context

In technical systems, bias is most commonly understood and treated as a statistical phenomenon. Bias is an effect that deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error, which may distort on any one occasion but balances out on the average [71]. The International Organization for Standardization (ISO) defines bias more generally as: "the degree to which a reference value deviates from the truth"[72]. In this context, an AI system is said to be biased when it exhibits systematically inaccurate behavior. This statistical perspective does not sufficiently encompass or

communicate the full spectrum of risks posed by bias in AI systems.

Legal context

This section was developed in response to public comments. Stakeholder feedback noted that the discussion of bias in AI could not be divorced from the treatment of bias in the U.S. legal system and how it relates to laws and regulations addressing discrimination and fairness, especially in the areas of consumer finance, housing, and employment.[6,7] There currently is no uniformly applied approach among the regulators and courts to measuring impermissible bias in all such areas. Impermissible discriminatory bias generally is defined by the courts as either consisting of disparate treatment, broadly defined as a decision that treats an individual less favorably than similarly situated individuals because of a protected characteristic such as race, sex, or other trait, or as disparate impact, which is broadly defined as a facially neutral policy or practice that disproportionately harms a group based on a protected trait.[8]

This section is presented not as legal guidance, rather as a reminder for developers, deployers, and users of AI that they must be cognizant of legal considerations in their work, particularly with regard to bias testing. This section provides basic background understanding of some of the many ways bias is treated in some federal laws.

As it relates to disparate impact, courts and regulators have utilized or considered as acceptable various statistical tests to evaluate evidence of disparate impact. Traditional methods of statistical bias testing look at differences in predictions across protected classes, such as race or sex. In particular, courts have looked to statistical significance testing to assess whether the challenged practice likely caused the disparity and was not the result of chance or a nondiscriminatory factor.[9]

[6]Many laws, at the federal, state and even municipal levels focus on preventing discrimination in a host of areas. *See e.g.* Title VII of the Civil Rights Act, regarding discrimination on the basis of sex, religion, race, color, or national origin in employment, the Equal Credit Opportunity Act, focused, broadly, on discrimination in finance, the Fair Housing Act, focused on discrimination in housing, and the Americans with Disabilities Act, focused on discrimination related to disabilities, among others. Other federal agencies, including the U.S. Equal Employment Opportunity Commission, the Federal Trade Commission, the U.S. Department of Justice, and the Office Federal Contract Compliance Programs are responsible for enforcement and interpretation of these laws.

[7]Note that the analysis in this section is not intended to serve as a fully comprehensive discussion of the law, how it has been interpreted by the courts, or how it is enforced by regulatory agencies, but rather to provide an initial high-level overview.

[8]*See* 42 U.S.C. 2000e-2(a) (2018) and 42 U.S.C. 2000e-2(k) (2018), respectively.

[9]The Uniform Guidelines on Employment Selection Procedures (UGESP) state "[a] selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5ths) (or eighty percent) of the rate for the group

It is important to note, however, that the tests used to measure bias are not applied uniformly within the legal context. In particular, federal circuit courts are split on whether to require a plaintiff to demonstrate both statistical and practical significance to make out a case of disparate impact. Some decisions have expressly rejected practical significance tests in recent years while others have continued to endorse their utility. This split illustrates that while the legal context provides several examples of how bias and fairness has been quantified and adjudicated over the last several decades, the relevant standards are still evolving.

It is also important to note that critical differences exist between traditional disparate impact analyses described above and illegal discrimination as it relates to people with disabilities, particularly under the Americans with Disabilities Act (ADA). Claims under the ADA are frequently construed as "screen out" rather than as "disparate impact" claims. "Screen out" may occur when an individual with a disability performs poorly on an evaluation or assessment, or is otherwise unable to meet an employer's job requirements, because of a disability and the individual loses a job opportunity as a result. In addition, the ADA's prohibition against denial of reasonable accommodation, for example, may require an employer to change processes or procedures to enable a particular individual with a disability to apply for a job, perform a job, or enjoy the benefits and privileges of employment. Such disability-related protections are particularly important to AI systems because testing an algorithm for bias by determining whether such groups perform equally well may fail to detect certain kinds of bias. Likewise, eliminating group discrepancies will not necessarily prevent screen out or the need for reasonable accommodation in such systems.

Cognitive and societal context

The teams involved in AI system design and development bring their cognitive biases, both individual and group, into the process [73]. Bias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed — or if AI is required at all. There are systemic biases at the institutional level that affect how organizations and teams are structured and who controls the decision making processes, and individual and group heuristics and cognitive/perceptual biases throughout the AI lifecycle (as described in Section 2.4). Decisions made by end users, downstream decision makers, and policy makers are also impacted by these biases, can reflect limited points of view and lead to biased outcomes [74–79]. Biases impacting human decision making are usually implicit and unconscious, and therefore unable to be easily controlled or mitigated [80]. Any assumption that biases can be remedied by human control or awareness is not a recipe for success.

with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." 29 C.F.R. § 1607.4(D)