NIST Special Publication 1270

# Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

Reva Schwartz
Apostol Vassilev
Kristen Greene
Lori Perine
Andrew Burt
Patrick Hall

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1270

National Institute of Standards and Technology
U.S. Department of Commerce

# NIST Special Publication 1270

# Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

Reva Schwartz
*National Institute of Standards and Technology*
*Information Technology Laboratory*

Apostol Vassilev
*National Institute of Standards and Technology*
*Information Technology Laboratory*
*Computer Security Division*

Kristen Greene
*National Institute of Standards and Technology*
*Information Technology Laboratory*
*Information Access Division*

Lori Perine
*National Institute of Standards and Technology*
*Information Technology Laboratory*
*& The University of Maryland*

Andrew Burt
Patrick Hall
*BNH.AI*

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1270

March 2022


U.S. Department of Commerce
*Gina M. Raimondo, Secretary*

National Institute of Standards and Technology
*James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology*

**National Institute of Standards and Technology Special Publication 1270**
**Natl. Inst. Stand. Technol. Spec. Publ. 1270, 86 pages (March 2022)**
**CODEN: NSPUE2**

**This publication is available free of charge from:**
**https://doi.org/10.6028/NIST.SP.1270**

## Executive Summary

As individuals and communities interact in and with an environment that is increasingly virtual, they are often vulnerable to the commodification of their digital footprint. Concepts and behavior that are ambiguous in nature are captured in this environment, quantified, and used to categorize, sort, recommend, or make decisions about people's lives. While many organizations seek to utilize this information in a responsible manner, biases remain endemic across technology processes and can lead to harmful impacts regardless of intent. These harmful outcomes, even if inadvertent, create significant challenges for cultivating public trust in artificial intelligence (AI).

While there are many approaches for ensuring the technology we use every day is safe and secure, there are factors specific to AI that require new perspectives. AI systems are often placed in contexts where they can have the most impact. Whether that impact is helpful or harmful is a fundamental question in the area of Trustworthy and Responsible AI. Harmful impacts stemming from AI are not just at the individual or enterprise level, but are able to ripple into the broader society. The scale of damage, and the speed at which it can be perpetrated by AI applications or through the extension of large machine learning MODELs across domains and industries requires concerted effort. Current attempts for addressing the harmful effects of AI bias remain focused on computational factors such as representativeness of datasets and fairness of machine learning algorithms. These remedies are vital for mitigating bias, and more work remains. Yet, as illustrated in Fig. 1, human and systemic institutional and societal factors are significant sources of AI bias as well, and are currently overlooked. Successfully meeting this challenge will require taking all forms of bias into account. This means expanding our perspective beyond the machine learning pipeline to recognize and investigate how this technology is both created within and impacts our society.

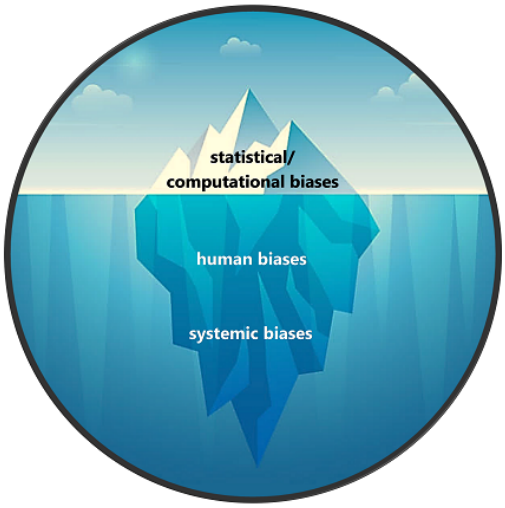


**Fig. 1.** The challenge of managing AI bias

Trustworthy and Responsible AI is not just about whether a given AI system is biased, fair or ethical, but whether it does what is claimed. Many practices exist for responsibly producing AI. The importance of transparency, datasets, and test, evaluation, validation, and verification (TEVV) cannot be overstated. Human factors such as participatory design techniques and multi-stakeholder approaches, and a human-in-the-loop are also important for mitigating risks related to AI bias. However none of these practices individually or in

concert are a panacea against bias and each brings its own set of pitfalls. What is missing from current remedies is guidance from a broader SOCIO-TECHNICAL perspective that connects these practices to societal values. Experts in the area of Trustworthy and Responsible AI counsel that to successfully manage the risks of AI bias we must operationalize these values and create new norms around how AI is built and deployed. This document, and work by the National Institute of Standards and Technology (NIST) in the area of AI bias, is based on a socio-technical perspective.

The intent of this document is to surface the salient issues in the challenging area of AI bias, and to provide a first step on the roadmap for developing detailed socio-technical guidance for identifying and managing AI bias. Specifically, this special publication:

- describes the stakes and challenge of bias in artificial intelligence and provides examples of how and why it can chip away at public trust;
- identifies three categories of bias in AI — systemic, statistical, and human — and describes how and where they contribute to harms;
- describes three broad challenges for mitigating bias — datasets, testing and evaluation, and human factors — and introduces preliminary guidance for addressing them.

Bias is neither new nor unique to AI and it is not possible to achieve zero risk of bias in an AI system. NIST intends to develop methods for increasing assurance, GOVERNANCE and practice improvements for identifying, understanding, measuring, managing, and reducing bias. To reach this goal, techniques are needed that are flexible, can be applied across contexts regardless of industry, and are easily communicated to different stakeholder groups. To contribute to the growth of this burgeoning topic area, NIST will continue its work in measuring and evaluating computational biases, and seeks to create a hub for evaluating socio-technical factors. This will include development of formal guidance and standards, supporting standards development activities such as workshops and public comment periods for draft documents, and ongoing discussion of these topics with the stakeholder community.

**Key words**

bias, trustworthiness, AI safety, AI lifecycle, AI development

**Acknowledgments**
The authors wish to thank everyone who responded to our call and submitted comments to the draft version of this paper. The received comments and suggested references were essential for improving the paper and the future direction of this work. We also want to thank the many people who assisted with the updating of the document, including our NIST colleagues, and other reviewers who took the time to provide their constructive feedback. We thank Kyle Fox for his insightful comments, discussions, and invaluable input.

**Audience**
The intended primary audience for this document includes individuals and groups who are responsible for designing, developing, deploying, evaluating, and governing AI systems. The document is informed and motivated by segments of the public who experience potential harm or inequities due to bias in AI systems, or are affected by biases that are newly introduced or amplified by AI systems.

**Background**
This document is a result of an extensive literature review, conversations with experts from the areas of AI bias, fairness, and socio-technical systems, a workshop on AI bias,[1] and public comments on the draft version.[2] Insights derived from the public comments have been integrated throughout this document. An overview and analysis of themes from the public comments will be posted.[3] Intermediate follow-on work to this publication will include development of formal guidance for assessing and managing the risks of AI bias, and a series of public workshops to discuss these topics with the stakeholder community and build consensus.

**Trademark Information**


**NIST Special Publications**
The National Institute of Standards and Technology (NIST) promotes U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life. Among its broad range of activities, NIST contributes to the research, standards, evaluations, and data required to advance the development, use, and assurance of trustworthy artificial intelligence (AI).

[1]For more information about this workshop see https://www.nist.gov/news-events/events/2020/08/bias-ai-workshop.
[2]Public comments are available at https://www.nist.gov/artificial-intelligence/comments-received-proposal-identifying-and-managing-bias-artificial.
[3]Updated information for all of these resources can be found on the NIST AI Bias webpage, located at https://www.nist.gov/artificial-intelligence/ai-fundamental-research-free-bias.

The Information Technology Laboratory (ITL) at NIST develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines.

This special publication focuses on addressing and managing risks associated with bias in the design, development, and use of AI. It is one of a series of documents and workshops related to the NIST AI Risk Management Framework (AI RMF) and is intended to advance the trustworthiness of AI technologies. As with other documents in the AI RMF series, this publication provides reference information and technical guidance on terminology, processes and procedures, and test and evaluation, validation, and verification (TEVV). While practical guidance[4] published by NIST may serve as an informative reference, this guidance remains voluntary.

The content of this document reflects recommended practices. This document is not intended to serve as or supersede existing regulations, laws, or other mandatory guidance.

[4]The term 'practice guide,' 'guide,' 'guidance' or the like, in the context of this paper, is a consensus-created, informative reference intended for voluntary use; it should not be interpreted as equal to the use of the term 'guidance' in a legal or regulatory context." This document does not establish any legal standard or any other legal requirement or defense under any law, nor have the force or effect of law.

**How to read this document**

Section 1 lays out the purpose and scope of NIST's work in AI bias. Section 2 describes three categories of bias and how they may occur in the commission, design, development, and deployment of AI technologies that can be used to generate predictions, recommendations, or decisions (such as the use of algorithmic decision systems), and how AI systems may impact individuals and communities or create broader societal harms. Section 3 describes the challenge of bias related to three core areas: datasets; test, evaluation, validation and verification; and human factors, and provides general guidance for managing AI bias in each of those areas.

This document uses terms such as AI technology, AI system, and AI applications interchangeably. Terms related to the machine learning pipeline, such as AI model or algorithm are also used in this document interchangeably. Depending on context, when the term "system" is used it may refer to the broader organizational and/or social ecosystem within which the technology was designed, developed, deployed, and used, instead of the more traditional use related to computational hardware or software.

Important reading notes:

- The document includes a series of vignettes, shown in red callout boxes, to help exemplify how and why AI bias can reduce public trust. Interesting nuances/aspects are highlighted in blue callout boxes, important takeaways are shown as framed text.
- Terms that are displayed as small caps in the text are defined in the GLOSSARY. Clicking on a word shown in small caps, e.g. MODEL, takes the reader directly to the definition of that term in the Glossary. From there, one may click on a page number shown at the end of the definition to return.
- March 24, 2022 update: the following changes are introduced with respect to the original version of this document published on March 15, 2022:
  - Fixed typos in the text of Fig. 5 and Fig. 7.
  - Removed duplicates and fixed poorly formatted entries in the **References**.
  - Corrected a statement in the text of VIGNETTE on p.7 regarding the work cited in [36].

# Contents



# List of Figures

## 1. Purpose and Scope

In August 2019, fulfilling an assignment in an Executive Order on AI,[5] NIST released "A Plan for Federal Engagement in Developing Technical Standards and Related Tools" [1]. Based on broad public and private sector input, this plan recommended a deeper, more consistent, and long-term engagement in AI standards "to help the United States to speed the pace of reliable, robust, and trustworthy AI technology development." NIST research in AI continues along this path to focus on how to measure, evaluate, and enhance the trustworthiness of AI systems and the responsible practices for designing, developing, and deploying such systems. Working with the AI community, NIST has identified the following technical and socio-technical characteristics needed to cultivate trust in AI systems: accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security resilience—and that harmful biases are mitigated or controlled.

While AI has significant potential as a transformative technology, it also poses inherent risks. Since trust and risk are closely related, NIST's work in the area of trustworthy and responsible AI centers around development of a voluntary Risk Management Framework (RMF). The unique challenges of AI require a deeper understanding of how AI risks differ from other domains. The NIST AI RMF is intended to address risks in the design, development, use, and evaluation of AI products, services, and systems for such tasks as recommendation, diagnosis, pattern recognition, and automated planning and decision-making. The framework is intended to enable the development and use of AI in ways that will increase trustworthiness, advance usefulness, and address potential harms. NIST is leveraging a multi-stakeholder approach to creating and maintaining actionable practice guides via the RMF that is broadly adoptable.

### AI risk management

AI risk management seeks to minimize anticipated and emergent negative impacts of AI systems, including threats to civil liberties and rights. One of those risks is bias. Bias exists in many forms, is omnipresent in society, and can become ingrained in the automated systems that help make decisions about our lives. While bias is not always a negative phenomenon, certain biases exhibited in AI models and systems can perpetuate and amplify negative impacts on individuals, organizations, and society. These biases can also indirectly reduce public trust in AI. There is no shortage of examples where bias in some aspect of AI technology and its use has caused harm and negatively impacted lives, such as in hiring, [2–7] health care, [8–17] and criminal justice [18–30]. Indeed, there are many instances in which the deployment of AI technologies have been accompanied by concerns about whether and how societal biases are being perpetuated or amplified [31–46].

### Public perspectives

Depending on the application, most Americans are likely to be unaware of when they are

[5] Exec. Order No. 13,859, 84 Fed. Reg. 3,967 (Feb. 11, 2019), https://www.federalregister.gov/documents/2019/02/14/2019-02544/maitaining-american-leadership-in-artificial-intelligence.

interacting with AI enabled technology [47]. However, there is a general view that there needs to be a "higher ethical standard" for AI than for other forms of technology [48]. This mainly stems from the perceptions and fears about loss of control and privacy [46, 49–51].

Bias is tightly associated with the concepts of transparency and fairness in society. For much of the public, the assumptions underlying algorithms are rarely transparent. The complex web of code and decisions that went into the design, development, and deployment of AI rarely is easily accessible or understandable to non-technical audiences. Nevertheless, many people are affected by—or their data is used as inputs for—AI technologies and systems without their consent, such as when they apply to college, [52] for a new apartment, [53] or search the internet. When individuals feel that they are not being fairly judged when applying for jobs [2–5, 7, 54–56] or loans [57–59] it can reduce public trust in AI technology [60, 61].

When an end user is presented with information online that stigmatizes them based on their race, age, or gender, or doesn't accurately perceive their identity, it causes harm [34, 36, 37, 41]. Consumers can be impacted by price gouging practices resulting from an AI application, even when it is not used to make decisions directly affecting that individual [43].

## 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 [62]. Machine learning (ML) refers more specifically to the "field of study that gives computers the ability to learn without being explicitly programmed," [63] 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 [64, 65]. 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 [66–69].

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 [70]. The International Organization for Standardization (ISO) defines bias more generally as: "the degree to which a reference value deviates from the truth"[71]. 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 [72]. 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 [73–78]. Biases impacting human decision making are usually implicit and unconscious, and therefore unable to be easily controlled or mitigated [79]. 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)

### 2.1.2 Categories of AI bias

Based on previous academic work to classify AI bias [80–90] and discussions with thought leaders in the field, it is possible to identify three dominant categories of AI bias. This three-way categorization helps to expand our understanding of AI bias beyond the computational realm. By defining and describing how systemic and human biases present within AI, we can build new approaches for analyzing, managing, and mitigating bias and begin to understand how these biases interact with each other. Correspondingly, Fig. 2 presents three categories of AI bias. Definitions for these terms are found in the GLOSSARY.[10] This list of biases, while not exhaustive, constitutes prominent risks and vulnerabilities to consider when designing, developing, deploying, evaluating, using, or auditing AI applications.

**Systemic**
Systemic biases result from procedures and practices of particular institutions that operate in ways which result in certain social groups being advantaged or favored and others being disadvantaged or devalued. This need not be the result of any conscious prejudice or discrimination but rather of the majority following existing rules or norms. Institutional racism and sexism are the most common examples [91]. Other systemic bias occurs when infrastructures for daily living are not developed using universal design principles, thus limiting or hindering accessibility for persons with disabilities. Systemic bias is also referred to as institutional or historical bias. These biases are present in the datasets used in AI, and the institutional norms, practices, and processes across the AI lifecycle and in broader culture and society. See VIGNETTE for more examples.

[10] Definitions for each category of bias were often selected based on either recently published papers on the topic, or seminal work within the domain the term is most associated with. When multiple definitions were identified, the most relevant definition was selected or adapted. The references provided are not intended to indicate specific endorsement or to assign originator credit.

**Systemic bias in gender identification**

Beyond personal identity, human faces encode a number of conspicuous traits such as nonverbal expression, indicators of sexual attraction and selection, and emotion. Facial recognition technology (FRT) is used in many types of applications including gender identification, which compares morphological distances between faces to classify human faces by gender. The degree of sexual dimorphism between men and women appears to vary with age and ethnic group. As a consequence, accuracy of FRT gender identification can vary with respect to the age and ethnic group [92]. Prepubescent male faces are frequently misclassified as female, and older female faces are progressively misclassified as male [92]. Studies have highlighted that human preferences for sexually dimorphic faces may be evolutionarily novel [93, 94]. One study found differing levels of facial sexual dimorphism in samples taken from countries located in Europe, South America, and Africa [95]. Buolamwini and Gebru examined the suitability of using skin types as a proxy for demographic classifications of ethnicity or race and found that skin type is not an adequate proxy for such classifications. Multiple ethnicities can be represented by a given skin type, and skin type can vary widely within a racial or ethnic category. For example, the skin types of individuals identifying as Black in the U.S. can represent many hues, which also can be represented in ethnic Hispanic, Asian, Pacific Islander and American indigenous groups. Moreover, racial and ethnic categories tend to vary across geographies and over time [36]. While training data based on a limited or non-representative sample of a group results in lower accuracy in categorizing members of that group, the degree of sexual monomorphism or dimorphism within that group also affects accuracy. Additional biases can occur due to a lack of awareness about the multiplicity of gender [96].

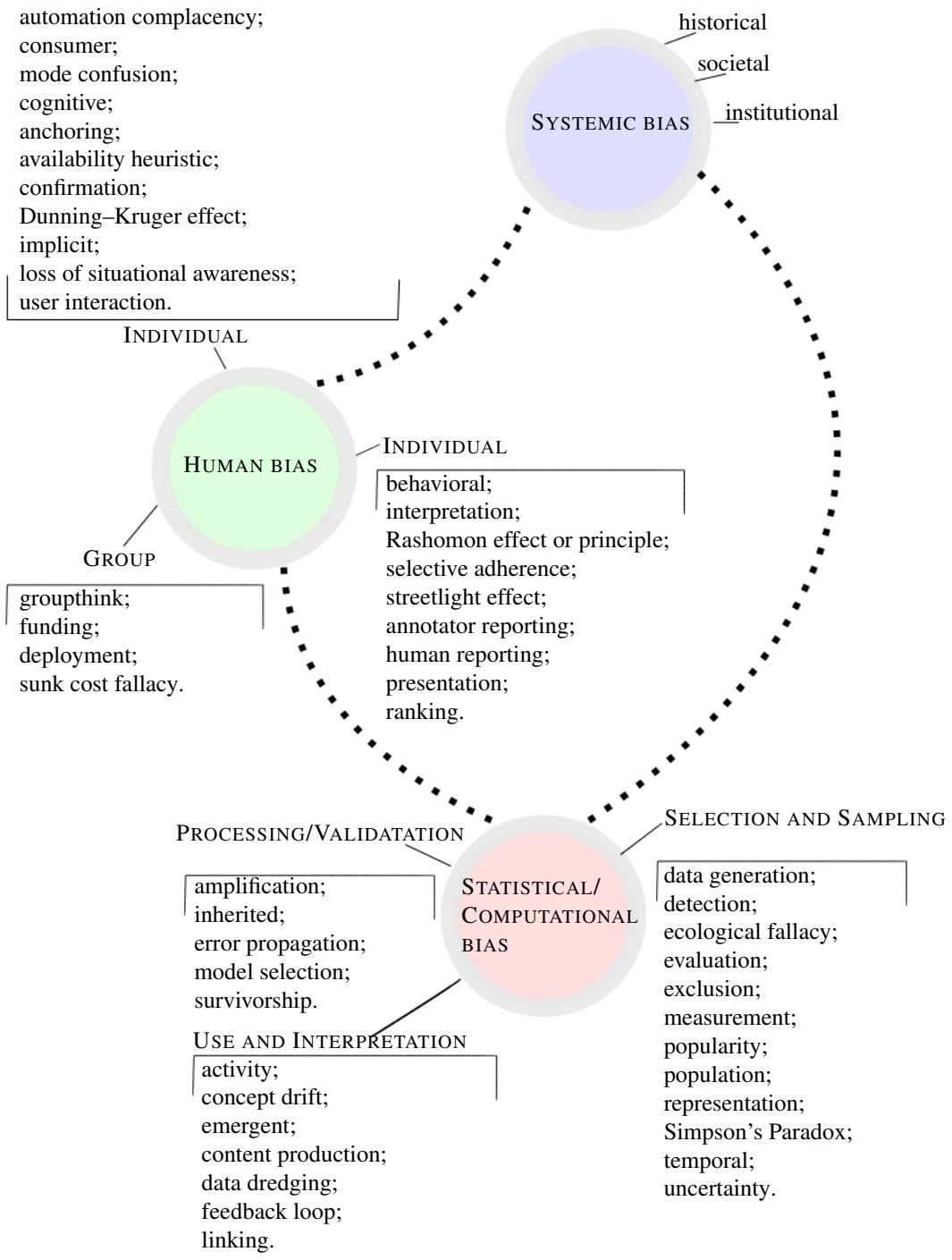


**Fig. 2.** Categories of AI Bias. The leaf node terms in each subcategory in the picture are hyperlinked to the GLOSSARY. Clicking them will bring up the definition in the Glossary. To return, click on the current page number (8) printed right after the glossary definition.

**Statistical and Computational**
Statistical and computational biases stem from errors that result when the sample is not representative of the population. These biases arise from systematic as opposed to random error and can occur in the absence of prejudice, partiality, or discriminatory intent [97]. In AI systems, these biases are present in the datasets and algorithmic processes used in the development of AI applications, and often arise when algorithms are trained on one type of data and cannot extrapolate beyond those data. The error may be due to heterogeneous data, representation of complex data in simpler mathematical representations, wrong data, and algorithmic biases such as over- and under-fitting, the treatment of outliers, and data cleaning and imputation factors.

**Human**
Human biases reflect systematic errors in human thought based on a limited number of heuristic principles and predicting values to simpler judgmental operations [98]. These biases are often implicit and tend to relate to how an individual or group perceives information (such as automated AI output) to make a decision or fill in missing or unknown information. These biases are omnipresent in the institutional, group, and individual decision making processes across the AI lifecycle, and in the use of AI applications once deployed. There is a wide variety of human biases. Cognitive and perceptual biases show themselves in all domains and are not unique to human interactions with AI. Rather, they are a fundamental part of the human mind. There is an entire field of study centered around biases and heuristics in thinking, decision-making, and behavioral economics for example [98]. Such research investigates phenomena such as ANCHORING BIAS, availability heuristic or bias, CONFIRMATION BIAS, and framing effects, among many others. It should be noted that heuristics are adaptive mental shortcuts that can be helpful, allowing complexity reduction in tasks of judgement and choice, yet can also lead to cognitive biases [98]. Human heuristics and biases are implicit; as such, simply increasing awareness of bias does not ensure control over it. Here we focus on broader examples of human bias in the AI space.

### 2.2 How AI bias contributes to harms

Technology based on AI has tighter connections to and broader impacts on society than traditional software. Applications that utilize AI are often deployed across sectors and contexts for decision-support and decision-making. In this role, they can replace humans and human processes for high-impact decisions. For example, AI-based hiring technologies and the models that underlie them replace people-oriented hiring processes and are implemented in any sector that seeks to automate their recruiting and employment pipeline [99–101]. Yet, ML models tend to exhibit "unexpectedly poor behavior when deployed in real world domains" without domain-specific constraints supplied by human operators [102]. These contradictions are a cause for considerable concern with large language models (or so-called foundation models) due to their considerable EPISTEMIC and ALEATORIC

uncertainty[103] (as described in Section 3.2.1)—among other factors. Methods for capturing the poor performance, harmful impacts and other results of these models currently are imprecise and non-comprehensive.

**Values**

While ML systems are able to model complex phenomena, whether they are capable of learning and operating in line with our societal values remains an area of considerable research and concern [55, 60, 104–109]. Systemic and implicit biases such as racism and other forms of discrimination can inadvertently manifest in AI through the data used in training, as well as through the institutional policies and practices underlying how AI is commissioned, developed, deployed, and used. Statistical/algorithmic and human cognitive and perceptual biases enter the engineering and modeling processes themselves, and an inability to properly validate model performance leaves these biases exposed during deployment [61, 102, 110, 111]. These biases collide with the cognitive biases of the individuals interacting with the AI systems as users, experts in the loop, or other decision makers. Teams that develop and deploy AI often have inaccurate expectations of how the technology will be used and what human oversight can accomplish, especially when deployed outside of its original intent [112, 113]. Left unaddressed, these biases and accompanying contextual factors can combine into a complex and pernicious mixture. These biases can negatively impact individuals and society by amplifying and reinforcing discrimination at a speed and scale far beyond the traditional discriminatory practices that can result from implicit human or institutional biases such as racism, sexism, ageism or ableism.

### 2.3 A Socio-technical Systems Approach

Likely due to expectations based on techno-solutionism and a lack of mature AI process governance, organizations often default to overly technical solutions for AI bias issues. Yet, these mathematical and computational approaches do not adequately capture the societal impact of AI systems [61, 73, 75, 111]. The limitations of a computational-only perspective for addressing bias have become evident as AI systems increasingly expand into our lives.

The reviewed literature suggests that the expansion of AI into many aspects of public life requires extending our view from a mainly technical perspective to one that is socio-technical in nature, and considers AI within the larger social system in which it operates [7, 19, 31, 37, 74, 75, 78, 114–119]. Using a socio-technical approach to AI bias makes it possible to evaluate dynamic systems of bias and understand how they impact each other and under what conditions these biases are attenuated or amplified. Adopting a socio-technical perspective can enable a broader understanding of AI impacts and the key decisions that happen throughout, and beyond, the AI lifecycle–such as whether technology is even a solution to a given task or problem [3, 108]. Reframing AI-related factors such as datasets, TEVV, participatory design, and human-in-the-loop practices through a socio-technical lens means understanding how they are both functions of society and, through the power of AI, can impact society. A socio-technical approach also enables analytic

approaches that take into account the needs of individuals, groups and society.

**Techno-solutionism**
As computational technologies have evolved, there has been an increasing tendency to believe that technical solutions alone are sufficient for addressing complex problems that may have social, political, ecological, economic, and/or ethical dimensions. This approach to problem-solving, often termed techno-solutionism,[120] assumes that the "right" code or algorithm can be applied to any problem and ignores or minimizes the relevance of human, organizational, and societal values and behaviors that inform design, deployment, and use of technology.

In the context of socio-technical AI systems, techno-solutionism promotes a viewpoint that is too narrow to effectively address bias risks. One control, for example, used in model risk management to mitigate against techno-solutionism and other anti-patterns, is to establish, document, and review the anticipated real-world value of an AI system.

Socio-technical approaches in AI are an emerging area, and identifying measurement techniques to take these factors into consideration will require a broad set of disciplines and stakeholders. Identifying contextual requirements for evaluating socio-technical systems is necessary. Developing scientifically supportable guidelines to meet socio-technical requirements will be a core focus.

AI bias extends beyond computational algorithms and models, and the datasets upon which they are built. The assumptions and decisions made within the processes used to develop technology are key factors, as well as how AI technology is used and interpreted once deployed. The idea that quantitative measures are better and more objective than other observations is known as the MCNAMARA FALLACY. This fallacy, and the related concept TECHNOCHAUVINISM [35], are at the center of many of the issues related to algorithmic bias. Traditional ML approaches attempt to turn ambiguity, context, human subjectivity, and categorical observations into objectively measurable quantities based on numerical mathematical models of their representations. This well-intentioned process enables data-driven modeling but it also inadvertently creates new challenges for socio-technical systems. Representing these complex human phenomena with mathematical models comes at the cost of disentangling the context necessary for understanding individual and societal impact and contributes to a fallacy of objectivity [121]. Science has made great strides in understanding the limitations of human cognition, including how humans perceive, learn, and store visual, aural, and textual information, and make decisions under risk. Yet, significant gaps remain. Thus, any mathematical attempt to model such human traits is limited and incomplete. This is a key challenge in model causality and predicting human interpretation of model output. And without proper governance, excising context and flattening the categories into numerical constructs makes traceability more difficult [122].

Finding approaches in TEVV to compensate for these limitations in the underlying modeling technology and bringing back the necessary context is an *important area of study.*

### 2.4 An Updated AI Lifecycle

Improving trust in AI by mitigating and managing bias starts with identifying a structure for how it presents within AI systems and uses. Organizations that design and develop AI technology use the AI lifecycle to keep track of their processes and ensure delivery of high-performing functional technology—but not necessarily to identify harms or manage them. This document has adapted a four-stage AI lifecycle from other stakeholder versions.[11] The intent is to enable AI designers, developers, evaluators and deployers to relate

[11] AI lifecycles utilized as key guidance in the development of the four-stage approach are: Centers of Excellence (CoE) at the U.S. General Services Administration [70] [IT Modernization CoE. (n.d.)], the Organisation for Economic Co-operation and Development [106] [Organisation for Economic Co-operation and Development. (2019).]. Another model of the AI lifecycle is currently under development with the Joint Technical Committee of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). *See* Information technology — Artificial intelligence — AI system life cycle processes, ISO/IEC CD 5338 (under development, 1st ed.), https://www.iso.org/standard/81118.html.

lifecycle processes with AI bias categories and effectively facilitate its identification and management. The academic literature and best practice guidelines strongly encourage a multi-stakeholder approach to developing AI applications using a lifecycle. Guidance for how organizations can enable this approach is described in Section 3.3.2 and focuses on participatory design methods such as human-centered design.

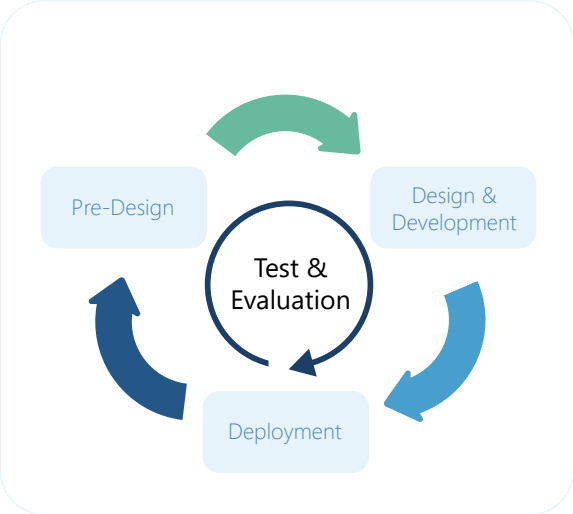


**Fig. 3.** The AI Development Lifecycle

AI Lifecycles are iterative, and begin in the **Pre-Design** stage, where planning, problem specification, background research, and identification of data take place. Decisions here include how to frame the problem, the purpose of the AI component, and the general notion that there is a problem requiring or benefiting from AI. Central to these decisions is who (individuals or groups) makes them and which individuals or teams have the most power or control over them. These early decisions and who makes them can reflect systemic biases within organizational settings, individual and group heuristics, and limited points of view. Systemic biases are also reflected in the datasets selected within pre-design. All of these biases can affect later stages and decisions in complex ways, and lead to biased outcomes [3, 74–78].

The **Design and Development** stage typically starts with analysis of the requirements and the available data. Based on this, a model is designed or selected. A compatibility analysis should be performed to ensure that potential sources of bias are identified and plans for mitigation are put into place. As model implementation progresses and is trained on selected data, the effectiveness of bias mitigation should be evaluated and adjusted.During development the organization should periodically assess the completeness of bias identification processes as well as the effectiveness of mitigation. Finally, at the end of the development stage, and before deployment, a thorough assessment of bias mitigation is necessary to ensure the system stays within pre-specified limits. The overall model specification must include the identified sources of bias, the implemented mitigation techniques and related performance assessments before the model can be released for deployment.

The **Deployment stage** is when the AI system is released and used. Once humans begin to interact with the AI system the performance of the system must be monitored and reassessed to ensure proper function. Teams should engage in continuous monitoring and have detailed policies and procedures for how to handle system output and behavior. System retraining may be necessary to correct adverse events, or decommission may be necessary. Since the lifecycle is iterative there are numerous opportunities for technology development teams to carry out multi-stakeholder consultation and ensure their applications

are not causing unintended effects or harms. Specific guidance for governing systems under these conditions is the subject of Section 3.4.1.

The **Test and Evaluation** stage is continuous throughout the entire AI Development Lifecycle. Organizations are encouraged to perform continuous testing and evaluation of all AI system components and features where bias can contribute to harmful impacts. For example, if during deployment the model is retrained with new data for a specific context, the model deployer should work with the model producer to assess actual performance for bias evaluation. Multi-stakeholder engagement is encouraged to ensure that the assessment is balanced and comprehensive. If deviations from desired goals are observed, the findings should feed into the model **Pre-Design** stage to ensure appropriate adjustments are made in data curation and problem formulation. Any proposed changes to the design of the model should then be evaluated together with the new data and requirements to ensure compatibility and identification of any potential new sources of bias. Then another round of design and implementation commences to formulate corresponding requirements for the new model capabilities and features and for additional datasets. During this stage, the model developer should perform continuous testing and evaluation to ensure that bias mitigation maintains effectiveness in the new setting, as the model is optimized and tested for performance. Once released, the deploying organization should use documented model specifications to test and evaluate bias characteristics during deployment in the specific context. Ideally, this evaluation should be performed together with other stakeholders to ensure all previously identified problems are resolved to everyone's satisfaction.

The most accurate model is not necessarily the one with the least harmful impact [123].

## 3. AI Bias: Challenges and Guidance

Through a review of the literature, and various multi-stakeholder processes, including public comments, workshops, and listening sessions, NIST has identified three broad areas that present challenges for addressing AI bias. The first challenge relates to **dataset** factors such as availability, representativeness, and baked-in societal biases. The second relates to issues of measurement and metrics to support testing and evaluation, validation, and verification (**TEVV**). The third area broadly comprises issues related to **human factors**, including societal and historic biases within individuals and organizations, as well as challenges related to implementing human-in-the-loop. This section outlines some key challenges associated with each of these three areas, along with recommended guidance.

It must be noted that TEVV does not amount to a full application of the scientific method. TEVV is an engineering construct that seeks to detect and remediate problems in a post-hoc fashion. The scientific method compels more holistic design thinking through

rigorous experimental design, hypothesis generation, and hypothesis testing. In particular, anecdotal evidence and the frequency of publicly-recorded AI bias incidents indicate that solid experimental design techniques that focus on structured data collection and selection and minimization of CONFIRMATION BIAS are being downplayed in many AI projects. CONSTRUCT VALIDITY is particularly important in AI system development. AI development teams should be able to demonstrate that the application is measuring the concept it intends to measure. It is important for all stakeholders, including AI development teams, to know how to evaluate scientific claims. That said, all the bias mitigants and governance processes outlined in this document do show promise. Interestingly, they are often borrowed from practices outside of core AI and ML — even technical guidance related to improved experimental design and more rigorous application of the scientific method. None are a panacea. All have pitfalls. NIST plans to work with the trustworthy and responsible AI communities to explore the proposed mitigants and governance processes, and build associated formal technical guidance over the coming years in concert with these communities.

The challenge of bias in AI is complex and multi-faceted. While there are many approaches for mitigating this challenge there is no quick fix.The recommendations in this document include a sampling of potentially promising techniques. These approaches, individually or in concert, are not a panacea against bias and each brings its own strengths and weaknesses.

## 3.1 Who is Counted? Datasets in AI Bias

### 3.1.1 Dataset Challenges

AI design and development practices rely on large scale datasets to drive ML processes. This ever-present need can lead researchers, developers, and practitioners to first "go where the data is," and adapt their questions accordingly [124]. This creates a culture focused more on which datasets are available or accessible, rather than what dataset might be most suitable [108]. As a result, the data used in these processes may not be fully representative of populations or the phenomena that are being modeled. The data that is collected can differ significantly from what occurs in the real world [76, 77, 117]. For example, sampling bias occurs when data is collected from responses to online questionnaires or is scraped from social media. The datasets which result are based on samples that are neither randomized nor representative of a population other than the users of a particular online platform. Such datasets are not generalizable, yet frequently are used to train ML applications which are deployed for use in broader socio-technical contexts, even though data representing certain societal groups may be excluded [116]. Systemic biases may also be manifested in the form of availability bias when datasets that are readily available but not

fully representative of the target population (including proxy data) are used and reused as training data. Disadvantaged groups including indigenous populations, women, and disabled people are consistently underrepresented [37, 116, 125, 126]. Similarly, datasets used in natural language processing (NLP) often differ significantly from their real-world applications, [127] which can lead to discrimination [128] and systematic gaps in performance. Other issues arise due to the common ML practice of reusing datasets. Under such practices, datasets may become disconnected from the social contexts and time periods of their creation. Scholars are beginning to examine the ethical and adverse impact implications of using data collected at a specific time for a specific purpose for uses that were not originally intended. Decontextualizing data raises questions related to privacy, consent, and internal validity of ML model results [129].

Even when datasets are representative, they may still exhibit entrenched historical and systemic biases, improperly utilize protected attributes, or utilize culturally or contextually unsuitable attributes. Developers sometimes exclude protected attributes, associated with social groups which have historically been discriminated against. However, this does not remedy the problem, since the information can be inadvertently inferred in other ways through proxy or latent variables. Latent variables such as gender can be inferred through browsing history, and race can be inferred through zip code. So models based on such variables can still negatively impact individuals or classes of individuals [73]. Thus, the proxies used in development may be both a poor fit for the concept or characteristic seeking to be measured, and reveal unintended information about persons and groups. There is also sensitivity related to attributes and inferences that do not receive protection under civil rights laws, but which may enable discrimination when inferred and used by an ML model, such as low income status. Alternately, when there is not sufficient knowledge or awareness of the socio-technical context of a process or phenomenon, the attributes that are collected for use in an ML application may not be universally applicable for modeling the different social groups or cultures who are analyzed using the application. For example, using (past) medical costs to predict the need for future health interventions leads to severe under-prediction of healthcare needs in groups that do not have sufficient access to health care, such as African Americans [14].

**Protected attributes:** A host of laws and regulations have been established to prohibit discrimination based on grounds such as race, sex, age, religious affiliation, national origin, and disability status, among others. Local laws can apply protections across a wide variety of groups and activities.

Once end users start to interact with an AI system, any early design and development decisions that were poorly or incompletely specified or based on narrow perspectives can be exposed, leaving the process vulnerable to additive statistical or human biases [77]. By not designing to compensate for activity biases, algorithmic models may be built on data from

only the most active users, likely creating downstream system activity that does not reflect the intended or real user population [130, 131] resulting in potentially harmful impacts. In one example, by considering that ads for jobs in Science, Technology, Engineering and Mathematics (STEM) might be seen most often by men due to how marketing algorithms optimize for cost in ad placement, the women who were the intended audience of the ads never saw them [132] cf., VIGNETTE for details. Furthermore, feedback loops can result in disparity amplification in which marginalized individuals or groups are less likely to use an AI system and the subsequent training data are based on the most frequent users. For example, non-native English speakers are less likely to use a voice-enabled personal assistant and people living in transit deserts are often dependent on ride-hailing services. So, the experiences of these groups do not match the intended purpose or operation of the AI system.

### 3.1.2 Dataset Guidance

A key question that must be asked for the development and deployment of an AI system is: *do datasets exist that are fit or suitable for the purpose of the various applications, domains and tasks for which the AI system is being developed and deployed?* Not only is the predictive behavior of the ML system determined by the data, but the data also largely defines the machine learning task itself [61]. The question of dataset fit or suitability requires attention to three factors: statistical methods for mitigating representation issues; processes to account for the socio-technical context in which the application is being deployed; and awareness of the interaction of human factors with the AI technical system at all stages of the AI lifecycle. When datasets are available, the set of metrics for demonstrating fairness are many, context-specific, and unable to be reduced to a concise mathematical definition [133].

**Statistical Factors** AI bias problems are exacerbated by the variety of statistical biases that are prevalent in the large scale datasets used in ML modeling. When these models are deployed for decision-based applications, often in high-risk settings and off-label uses, harms can be perpetuated and amplified.

A major trend for addressing AI bias is to focus on balanced statistical representation in the datasets used in modeling processes. Simple but effective techniques, such as class imbalance measures or label imbalance measures, or analysis using statistical phenomena such as SIMPSON'S PARADOX,[134] can be used to detect bias in datasets, and sometimes help mitigate it [85, 135–138]. Numerous studies and software libraries invoke data rebalancing processes (e.g., [139]). Causal models and graphs may also be used to detect direct discrimination in the data [61, 85].

Generalized linear models require that variables are independent with little multicollinearity and that residuals are normally distributed and homoscedastic. Furthermore, common algorithmic techniques such as $L^1$ and $L^2$ regularization in ML cost functions assume that the variables are unimodal. However, data is often heterogeneous and multimodal espe-

cially when populations are not disaggregated by gender, age, race, or income.

Thus, it is important to document and communicate the limitations of the applicability of AI outputs, whether a model is used for benchmarking, prediction, or classification. In many cases, practitioners train models on benchmark datasets and use them on real data in specific applications. However, it may not be possible to fully address mathematically the imbalances in representation and the heterogeneous nature of real-world heterogeneous datasets. A recent study highlighted serious errors in commonly used benchmark dataset [140]. Consequently, a model trained on biased and erroneous data may lead to biased and inaccurate predictions. Moreover, training a model on one dataset and using it to operate on another requires special care to account for potential differences in the distributions of the datasets that may further exacerbate the unfairness and errors of the model.

**Accounting for Socio-technical Factors**

While statistical methods are indeed necessary, they are not sufficient for addressing the AI bias challenges associated with datasets. Modeling processes have the intent of making contextual concepts measurable. Once the context has been removed, however, it is difficult to get it back, leading AI models to learn from inexact representations. Just as building codes are designed based on general principles, but designed to incorporate the specific geographic characteristics of a region, so too must the use of datasets in ML applications be adapted to take into the full spectrum of socio-technical factors of the context in which they are deployed.

> Word embeddings represent text data as positions in a high-dimensional mathematical space. Such a representation allows arithmetic (measurable comparisons) to be performed on words [141]. However, when text data are simplified as mathematical objects, contextual information including homographs or idioms that do not fit neatly into the model may be lost. When asked to compute "doctor" - "father" + "mother" using this arithmetic, an AI system might respond with "nurse." Is the AI system's answer due to historical gender stereotypes in professions or due to the natural, close association of the gender-specific verb "nurse" with mother? In other scenarios, even when attempts are made to explicitly remove bias from training data, biases may still exist because of deep, complex connections within the text data [80, 142].

Attention to the socio-technical factors for an AI system is essential at all phases of the lifecycle, most importantly in design, development, and deployment. In the design phase, socio-technical analysis provides insights into social variations in the dynamics or characteristics of a phenomenon. This can help better frame questions for analysis and enable assessment of dataset fit. A socio-technical perspective in the development phase facilitates selection of data sources and attributes, and explicitly integrates impact assessment as a complement to algorithmic accuracy. Studies have shown how it is possible to mathematically address statistical bias in a dataset, then develop an algorithm which performs with

high accuracy, yet produce outcomes that are harmful to a social class and diametrically opposed to the intended purpose of the AI system [14]. The need for new ways to measure the impact of AI systems is a current theme in the literature and the trustworthy and responsible AI research community. The practice of deploying AI in off-label uses, that is AI systems being applied to a task or within a social or organizational context for which it was not designed, must be approached with caution, especially in high-risk settings. Socio-technical analysis can help determine if such use, with modification, is both ethically and technically feasible. In all cases, a socio-technical perspective implicates adopting processes that include involving stakeholders, examining cultural dynamics and norms, and assessing societal impacts.

AI technologies can be perfectly accurate and still contribute to harmful outcomes.

**Interaction of human factors and datasets** Systemic institutional biases are captured in the datasets used to build the models underlying AI applications. These biases are compounded by the decisions and assumptions made by AI design and development teams about which datasets to use [129]. These decisions affect who and what gets counted, and who and what does not get counted. The issue of "flattening" the societal and behavioral factors within the datasets themselves is problematic, but often overlooked [66, 129, 143, 144]. The problem is further exacerbated by the variety of statistical biases that are prevalent in the large scale datasets used in ML modeling.

Human biases, whether conditioned socially or unconscious cognitive bias, are factors in data selection, curation, preparation and analysis processes. A person who annotates training data (for example, for gesture recognition and sentiment analysis) may impart their own perception biases. A person who chooses which data sources and variables to leave in or take out may do so in a way that aligns with a held belief. Data typically needs to be cleaned in some way, removing outliers and spurious data. Missing data may be imputed (replacing the missing values with nearest neighbors or extrapolated values) or removed entirely. Missing data may be more frequent in marginalized populations. Furthermore, because of compounding collection biases, missing and spurious data is often not random. Data analysis decisions such as the cardinal treatment of ordinal data in a Likert-scale or rating-scale data may lead to a biased estimator [145]. Processes for documenting potential sources of human bias are essential but often overlooked elements for characterizing AI model transparency and explainability, in addition to addressing AI bias and fairness. As with statistical factors and socio-technical analysis, incorporating awareness and documentation in the AI lifecycle helps to define limitations and ensure ethically and socially appropriate uses that do not perpetuate or amplify harms. See Section 3.3 for a more thorough discussion of challenges and guidance related to human factors and AI bias.

### 3.2 How do we know what is right? TEVV Considerations for AI Bias

#### 3.2.1 TEVV Challenges

Delegating decision-making to algorithms is appealing because ML systems produce more consistent decisions compared to humans [146]. However, AI systems do not work in a vacuum. Operational context, such as the jurisdiction and industry vertical in which a system operates, serves to frame fairness goals. Even the algorithm itself relies on data for training and performance tuning, which in turn can be assessed by a fairness metric. Therefore, when we consider the computational approaches to mitigating bias, we must take into consideration these three components together: algorithms, data, and fairness metrics.

AI systems regularly model concepts that are—at best—only partially observable or capturable by data. Without direct measures for these highly complex considerations, AI development teams use proxies, which can create many risks [147]. For example, for "criminality," a measurable index or construct, might be created from other information, such as arrests and convictions, which are used as PROXY variables for predicting a certain outcome—in this case, whether a certain individual is likely to be a repeat offender. In algorithmic hiring, an AI system might be developed using input variables such as "length of time in prior employment," "productivity," and "number of lost hours" as measurable proxies in lieu of the not directly measurable concept of "employment suitability." The algorithm might also include a predictor variable such as distance from the employment site [148] because it might correlate with employees quitting their job due to long commutes or bad traffic. However, since "distance from the employment site" might disadvantage candidates from certain neighborhoods, and "length of time in prior employment" might disadvantage candidates who are unable to find stable transportation (or relate to other socio-economic factors) the AI system will contribute to biased outcomes.

**Epistemic and aleatoric uncertainty**

ML distinguishes two types of predictive uncertainty: EPISTEMIC and ALEATORIC [149]. For example, models produced by deep learning ML systems exhibit epistemic uncertainty in the parameters of the computed model. The model parameters are typically computed as the result of a nonconvex minimization of an appropriately chosen cost function. It is well known from mathematics that such a formulation of the problem does not have a unique solution [150, 151]. While epistemic uncertainty can be reduced by increasing the amount of representative training data, it cannot be fully eliminated. This can impact the behavior of a deep learning system in deployment when used with real-world data, especially when there is a mismatch in the distributions of the real and training data [102]. This can lead to undesirable effects on many of the AI system's critical attributes (e.g., robustness, resilience), including inducing harmful bias. Even convex problems (e.g., multiple linear regression) may suffer from epistemic uncertainty when a decision variable is not included in the model.

Another inherent type of uncertainty associated with machine learning is ALEATORIC.

It represents the uncertainty inherent in the data, e.g., the uncertainty in the label assigning process of the training dataset. Aleatoric uncertainty is the irreducible part of the predictive uncertainty. Since these two types of uncertainties (EPISTEMIC and ALEATORIC) are highly context-dependent, changing the context may blur the difference between them or even cause one to turn into the other. Thus, their characterization as reducible and irreducible is not absolute. For example, datasets containing overlapping samples with different attributes could be embedded into higher dimensions so that the samples are clearly separated, thus reducing aleatoric uncertainty at the expense of epistemic uncertainty - because the model would likely overfit the existing data in the larger space. Some of the difficulty in distinguishing epistemic and aleatoric uncertainty is that ML models are (implicit) mathematical representations of the data on which they are trained [152].

**The growth of Large Language Models**

Large LANGUAGE MODELS (LLMs) have become the dominant trend in deep learning today and are expected to continue to grow in importance [103, 153]. Although LLMs have been able to achieve impressive advances in performance on a number of important tasks, they come with significant risks that could potentially undermine public trust in the technology. LLMs create significant challenges for both EPISTEMIC and ALEATORIC uncertainty. Relying on large amounts of uncurated web data increases aleatoric uncertainty [154]. In-depth knowledge of the data and its statistical properties is critically important for detecting bias in the predictive output of ML models.

Identifying sources of bias is the first step in any bias mitigation strategy.

**Epistemic uncertainty and large-scale AI models**
With the availability of large and fast computing resources, massive artificial neural networks are becoming increasingly common. In particular, some language models now consist of trillion-dimensional parameter spaces trained on hundreds of gigabytes of data. The training data, often scraped from internet sources, commonly has known gender, racial, cultural, and socio-economic biases [154, 155]. Alternative approaches to large-sized language datasets have been proposed to mitigate harmful bias, but such an approach may introduce other human biases in the selection of values-targeted datasets. Beyond the systemic and selection biases, large language models also highlight EPISTEMIC UNCERTAINTY. Stochastic gradient descent (or other accelerated methods) methods [151] are used to find a set of parameters that minimize a cost function associated with the model, but deep neural networks exhibit complicated nonlinearities which result in many potential local minima. A trillion-dimensional manifold may have a huge, unknown number of minima [156]. Furthermore, to fit these parameters into computer memory, it is often necessary to use half-precision floating-point numbers [157], introducing rounding error which may undermine stability in the numerical methods [158]. As a result, the model may demonstrate unknown and erratic behavior and challenges for reproducibility and explainability [159].

In the quest for fitting larger and larger models into existing finite computational resources, LLMs rely on techniques, e.g., reduced-precision numerical representations of models, that further increase the epistemic uncertainty of deep learning models, [160] cf., VIGNETTE. Early practice has shown that concerns about the use of LLMs are indeed valid, with preliminary experimental results showing LLMs exhibit significant bias [154, 161, 162]. To reduce risks from the use of LLMs, future work in this area should move towards efforts to fully understand and characterize their behavior, and to devise effective mitigation measures against the biases they bring.

**Processes**
While datasets exhibit numerous biases that lead to harmful impacts, they feed directly into other system level processes that determine what is important to model. For AI systems to determine this importance, and effectively categorize and sort the firehose of data for downstream recommendations and decisions, contextual information is flattened and unobservable phenomena are quantified through the development of indices and use of proxies. The use of data attributes with names like "criminality," "hireability," "creditworthiness," or similar can be indicative of experimental design problems that give rise to harmful bias.

The software designers and data scientists working in design and development are often highly focused on system performance and optimization. This focus can inadvertently be a source of bias in AI systems. For example, during model development and selection, modelers will almost always select the most accurate models. Yet, as Forde et al describe in their paper, [163] selecting models based solely on accuracy is not necessarily the best

approach for bias reduction. Furthermore, the choice of the model's objective function, upon which a model's definition of accuracy is based, can reflect bias. Not taking context into consideration during model selection can lead to biased results for sub-populations (for example, disparities in health care delivery). Relatedly, systems that are designed to use aggregated data about groups to make predictions about individual behavior—a practice initially meant to be a remedy for non-representative datasets[18]—can lead to biased outcomes. This bias, known as ECOLOGICAL FALLACY, occurs when an inference is made about an individual based on their membership within a group (for example, predicting college performance risk based on an individual's race [52]). These unintentional weightings of certain factors can cause algorithmic results that exacerbate and reinforce societal inequities.

Natural language processing (NLP) is a powerful computational approach to allow machines to meaningfully understand human spoken and written languages. Powering activities such as algorithmic search, speech translation, and even conversational text generation, NLP is able to help us communicate with computer systems to carry out a variety of tasks. The set of harms that can arise from the use of NLP however has become a recent concern in the area of trustworthy AI [80, 90, 154, 164, 165]. Hovy and Prabhumoye describe five sources of bias in NLP and potential ways to counteract it [166].

**Spurious Correlations**

The speed and scope of machine learning processes can unfortunately expand the development of systems based on questionable scientific underpinnings that learn spurious correlations related to human characteristics. For example, the German public radio outlet BR24 examined a system that purportedly assessed tone of voice, language, gestures, and facial expressions to create a personality profile for use in hiring processes [6]. The analysis showed the AI system was easily manipulated by superficial changes to its inputs, awarding candidates higher scores when they wore glasses or when a bookshelf was in the background, diminishing claims that the system analyzed human expressions, and raising concerns about shortcut learning [167]. Indeed, many AI systems now attempt to make inferences about individuals based on their facial characteristics that are not scientifically supportable, such as their propensity for committing crimes or even their sexual orientation [121, 168–172]. The basis for drawing conclusions about emotional state from facial characteristics ranges from unscientific and debunked theories to emerging experimental studies [173], presenting concerning challenges to AI systems that claim to make such judgements. By mechanizing human characteristics these systems can obfuscate significant uncertainty and result in harmful biases. AI-based hiring systems that claim to glean information about candidates from audio and video have been shown to increase bias in outcome decisions and may present untenable trade-offs between bias mitigation and prediction accuracy [174]. AI systems marketed as making predictions based on facial expressions often generate decisions based on biased experimental design premises [168] or spurious patterns learned by the system (e.g., shortcut learning). These cases illustrate the risks associated with using AI systems for tasks like sentiment or affect analysis, along with using systems to infer spurious correlations more broadly, which can perpetuate biases across groups and, in several instances can be scientifically unsound [175]. AI systems in consequential or sensitive areas should not be built on the basis of spurious correlations. They can provide faux-objective justification for biased outcomes. A socio-technical perspective broadens awareness of these risky computational approaches.

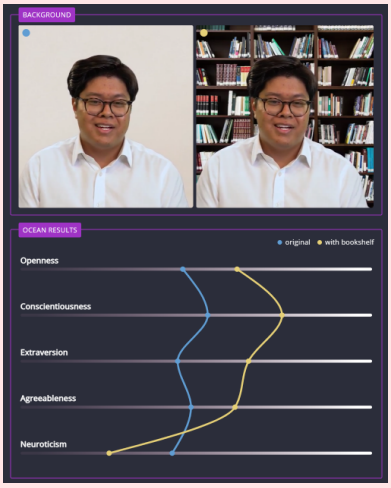


**Fig. 4.** The output of an AI system altered by background content.

The rise of predictive analytics as a mechanism for identifying patterns in human behavior is a recent example of a process that can produce biased outcomes and therefore

should be used carefully. These applications can be highly effective at identifying key insights in data that are unable to be gleaned by humans [176]. This technology is also often presented and perceived as a way to reduce human cognitive biases and make decisions more fair and objective [27, 177, 178]. In well defined and constrained settings these technologies can result in accurate and fair outcomes. However, the assumption that AI-based systems are more objective, especially in high stakes decision making, remains unclear. Categorizing unobservable behavior and phenomena leads to increased uncertainty in system performance. Measuring whether the patterns identified by these applications are real or a result of spurious correlations is difficult. Adding to the challenge is the reality that these systems are built and placed within organizational settings along with their accompanying — often unstated — policies and priorities, and used by subject matter experts and decision makers who have their own implicit heuristics and biases [179]. A fallacy of objectivity can often surround these processes, and may create conditions where technology's capacity and capabilities are oversold [121]. See VIGNETTE for an example.

**Algorithmic effects**

Algorithmic complexity can vary greatly across AI models. The number of parameters, which mathematically encode the training data, may be as few as one and as many as one trillion. Simple models with fewer parameters are often used because they tend to be less expensive to build, more explainable and more transparent, and easier to implement. However, such models can exacerbate statistical biases because restrictive assumptions on the training data often do not hold with nuanced demographics. Furthermore, designers who must make decisions on what variables to include or exclude can impart their own cognitive biases into the model [110, 180]. Complex models are often used on nonlinear, multimodal data such as text and images. Such models may capture latent systemic bias in ways that are difficult to recognize and predict. Expert systems, another AI paradigm, may encode cognitive and perceptual biases in the knowledge accumulated by practitioners from which the system is designed to emulate.

**Validity**

Ultimately, AI systems should demonstrate that they perform accurately, but how do we know what constitutes a "right answer"? Validating performance is a difficult but necessary endeavor for any system being deployed to the public and effective management and mitigation of AI bias. Many difficulties and flaws can arise in system validation. A common challenge in system testing is a lack of ground truth, or noisy labeling and other annotation factors which make it difficult to know what is accurate. The use of proxy variables compounds this difficulty, since what is being measured isn't directly observable. Performing system tests under optimal conditions — or conditions that are not close to the deployed state — is another challenging design flaw. System performance metrics are also difficult to generalize and can lead to issues with unintended use. Due to these challenges, subject matter experts should be relied upon during validation to create and oversee the most realistic possible validation processes [102]. Also the practice of "stratified perfor-

mance evaluations," [102] where system performance is analyzed across segments in the training or test data, whether demographic segments or otherwise, is a basic consideration for understanding system validity across a population of users.

**Validation and deployment**

Validation also means ensuring that the system is not being used in unintended ways. DEPLOYMENT BIAS happens when an AI model is used in ways not intended by developers. Emergent bias happens where the model is used in unanticipated contexts. Developers of an algorithm used by major U.S. cities to assist in coordinating housing to homeless people began phasing it out after several cities inappropriately used the algorithm as an assessment tool rather than as the presecreening tool as it was designed [181]. In another instance, the Chicago Police Department decommissioned an algorithm designed to predict the risk that an individual might be involved in future gun violence, citing unintended use and misapplication of the model [182].

It is not uncommon for deployment to be used as system testing. Depending on the context, institutional review may not be required to carry out this type of testing [183]. Without system validation, an AI system could be released that is technically flawed or fails to establish appropriate underlying mechanisms for proper functioning [184–186]. A system could be deployed in a negligent manner, be based on pseudoscience or spurious correlations, prey on the user, or generally exaggerate claims. In such cases, the goal should not be to ensure applications are bias-free, but to reject the development outright in order to prevent disappointment or harm to the user as well as to the reputation of the provider. Such systems may also run afoul of existing legal frameworks that proscribe unfair, deceptive, and predatory practices (UDAP).[12] This type of scenario may reinforce public distrust of AI technology since untested or technically flawed systems can contribute to bias and other harmful outcomes.

**AI systems as magic**

A further validation challenge of AI systems stems from their accessibility and hype. Physicist Richard Feynman referred to practices that superficially resemble science but do not follow the scientific method as cargo cult science. A core tenet of the scientific method is that hypotheses should be testable, experiments should be interpretable, and models should be falsifiable or at least verifiable. Commentators have drawn similarities between AI and cargo cult science citing its black box interpretability, reproducibility problem, and trial-and-error processes [187, 188]. High-level machine learning libraries and reduced costs of cloud computing have made AI more affordable and easier to develop. As a result, AI development is becoming increasingly democratized. Still, AI itself remains largely opaque—deep neural networks and Bayesian inference require advanced mathematics to understand. The DUNNING–KRUGER EFFECT is a cognitive bias in which a person with limited knowledge in a domain may vastly overestimate their understanding of that domain.

[12] *See, e.g.,* Federal Trade Commission Act, Section 5.

Even among experts, data-driven technologies can exacerbate CONFIRMATION BIAS, particularly when they are implicitly guided by expected outcomes. An analysis that examined hundreds of AI algorithms for identifying COVID found that few of them were effective [189].

The danger is that with enough tweaking of hyperparameters across many candidate AI models, one of them may appear to be highly accurate even when measured against standard performance datasets. DATA DREDGING (also known as p-hacking) is a statistical bias in which testing huge numbers of hypotheses of a dataset may appear to yield statistical significance even when the results are statistically nonsignificant.

Fig. 5 provides examples of how the three categories of bias — systemic, statistical and computational, and human - interact and contribute to harms within the data and processes used in AI applications, and the validation procedures for determining performance.

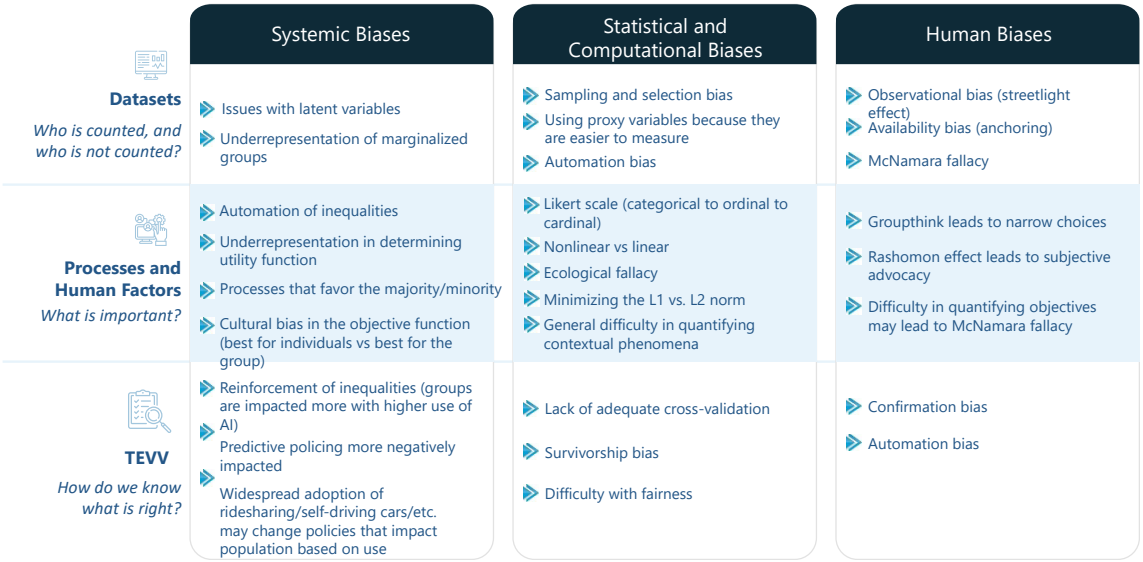


**Fig. 5.** How biases contribute to harms

### 3.2.2 TEVV Guidance

To mitigate the risks stemming from epistemic and aleatoric uncertainties, model developers should work closely with the organizations deploying them. Teams should work to ensure periodic model updates, and test and recalibrate model parameters on updated representative datasets to meet the business objectives while staying within desired performance targets and acceptable levels of bias. From a Bayesian inference perspective, this can be seen as updating the prior of the model to help avoid issues that may arise from using stale priors. Organizations are recommended to employ appropriate governance procedures to

adequately capture this cross-organizational need and ensure no negative impacts from using the AI technology.

**Algorithms**

In ML, it is not meaningful to assign bias to the model or algorithm itself without contextual information about the specific tasks on which they may be used. This links the model and algorithm to the dataset on which they are trained and tested (see VIGNETTE for how contextual factors can play a role in bias). The catchphrase "bias in, bias out" is widely used to describe the heavy dependence of the algorithmic behavior on the data. For example, in a natural language processing context, hate speech detection models use dialect markers as toxicity predictors, which can result in bias against minority groups [190]. In another context, an algorithm designed to deliver gender-neutral advertisements about jobs in STEM resulted in gender bias due to younger women being considered a valuable subgroup and more expensive as the targets for advertisements [85, 132].

Methods that help to reduce algorithmic bias are another helpful construct for understanding it. Specific methods for algorithmic mitigation of bias for many different machine learning tasks have been delineated or surveyed in recent studies [85, 191–194]. When considering approaches to mitigating algorithmic bias in a specific task context, recent literature categorizes debiasing methods into one of three categories [61, 85, 191, 194]:

1. **Pre-processing**: transforming the data so that the underlying discrimination is mitigated. This method can be used if a modeling pipeline is allowed to modify the training data.

2. **In-processing**: techniques that modify the algorithms in order to mitigate bias during model training. Model training processes could incorporate changes to the objective (cost) function or impose a new optimization constraint.

3. **Post-processing**: typically performed with the help of a holdout dataset (data not used in the training of the model). Here, the learned model is treated as a black box and its predictions are altered by a function during the post-processing phase. The function is deduced from the performance of the black box model on the holdout dataset. This technique may be useful in adapting a pre-trained large language model to a dataset and task of interest.

**The limits of algorithmic transparency in eliminating bias**

Automated decision-making is appealing but comes with risks that can result in discriminatory outcomes. Researchers investigated settings where ads are allocated by algorithm and found instances where historically–discriminated–against–groups are less likely to see desirable ads [132]. In this setting, a field test was performed with an ad that was intended to promote job opportunities and training in STEM. The STEM career ad campaign was motivated by widespread concern about a shortage of underrepresented groups in the STEM sector, particularly women. The assumption is that disseminating information about STEM careers to women and encouraging women to enter this field helps to address this problem. However, since women are far more likely to make decisions about household purchases, they are more valuable targets for advertising, creating pricing differentials for ad displays. The result of the ad campaign was that 20%+ more men than women viewed the ad, with the largest difference in the 25-54 year old age group.

The findings in this study help demonstrate the difficulty of evaluating algorithms for preventing discrimination, and the need for a socio-technical lens on the challenge. It is insufficient to look for bias in the algorithm alone. Relatedly, according to Lambrecht [132]:

*"One popular policy prescription has been a focus on algorithmic transparency where algorithmic codes are made public. Such policies are gaining increasing momentum - for example, the Federal Trade Commission (FTC) launched a new unit focused on algorithmic transparency, ... however, that algorithmic transparency would not have helped regulators to foresee uneven outcomes. The reason is that an examination of the algorithmic code would likely have revealed an algorithm focused on minimizing ad costs for advertisers. Without appropriate knowledge about the economic context and how such cost–minimization might affect the distribution of advertising, such 'transparency' would not have been particularly helpful."*

While transparency into AI system mechanisms is rarely a direct bias mitigant, as explained above, transparency enables many critical AI governance functions. Transparency is very important, but should not be mistaken for fairness.

In sectors of the U.S. economy where the Equal Credit Opportunity Act,[13] influential court cases,[14] or other legal and regulatory matters invoke the legal doctrine of Disparate Treatment, debiasing efforts may be less likely to explicitly include pre-, in-, and post-processing approaches, and instead rely on alternative modeling approaches. In consumer

[13] CFPB Supervision & Examination Manual, pt. II, § C, Equal Credit Opportunity Act (Oct. 2015).

[14] *e.g.*, Ricci v. DeStefano, 557 U.S. 557 (2009).

finance and employment litigation, where the practice of bias remediation, e.g., debiasing, has been pursued for decades, practitioners are more likely to consider adjustments to input variables or model hyperparameters to improve bias testing results or real-world outcomes. Demographic group membership, necessary for bias testing purposes, is often inferred using the Bayesian improved surname geocoding (BISG) process (see [195]).

Modeling algorithms or debiasing techniques that rely on demographic information, as most pre-, in-, and post-processing methods do, may pose higher risks in regulated environments where disparate treatment must be avoided [196].

**Fairness metrics**

From a computational standpoint, defining a fairness metric for ML requires developing a formal mathematical model to achieve desired predictive goals on a given dataset and associated task. Numerous fairness metrics are proposed in the literature [85, 191, 194, 197–199]. Much of the work in determining fairness criteria involves supervised learning, but the labeled data required for these tasks may not be readily available. This is particularly true for large language models, where the sheer scale of the datasets used for training is prohibitive for proper data labeling. This has a direct impact on both representativeness of the training data and, in turn, its impact on the representativeness of the generated model might exacerbate discriminatory outcomes, as large language models are adapted to specific datasets and tasks. Moreover, even if datasets are representative they may still exhibit biases or improperly utilize protected attributes, which in turn may lead to discrimination. Proxies may be used for hiding protected attributes and care should be taken to avoid discrimination resulting from badly chosen proxies [59, 136, 147, 200, 201]. And, even if proxies are used to hide protected attributes, they may still reveal sensitive information about individuals or groups [195, 202].

Recent literature [203] considers alternative learning tasks, e.g. unsupervised learning and reinforcement learning where only intermediate feedback is provided to the model, and tries to balance the effects of short- and long-term rewards. Several open questions still remain about the use and representativeness of synthetically generated data, in applications where little data is available. An emerging related line of research is to use simulations to evaluate the long-term impact of machine learning systems by incorporating elements of system level dynamics, feedback loops, and other long-term effects to make fair decisions in dynamic environments [204].

Another challenge, with serious social ramifications, is how to measure fairness in the emergent class of deployed generative models, such as large language models, computer vision systems, or deep fakes, whose outputs are free form text, audio or video [205].

While academic research into mathematical notions of fairness has blossomed in recent years, procedures for testing fairness in regulatory and litigation settings such as employment and consumer finance have been operational for decades, and reached a level of maturity before the recent increase in interest on the topic. In these areas, statistical tests can be applied to determine whether some automated decision-making system is acting outside the bounds of applicable law. $t$-tests, $\chi^2$-tests, analysis of regression coefficients, and other traditional statistical tests can be used to show a statistically significant difference between ML system outcomes across demographic groups. In some cases, measurements of differential validity are also used to ensure that applicants and employees receive roughly equal service from systems in employment, where system performance quality is evaluated across demographic groups.[15]

Credible attempts at bias mitigation should maintain alignment with acknowledged legal standards.

Generally, the majority of fairness metrics are observational as they can be expressed using probability statements involving the available random variables [61]. These metrics can be classified into many categories: fairness through unawareness, individual fairness, demographic parity, disparate impact, differential validity, proxy discrimination, equality of opportunity, etc. However, not all critically important lines of inquiry can be answered through observations alone. Moreover, depending on the relationship between a protected attribute and the data, certain observational definitions of fairness can increase discrimination. Hence, research to improve fairness metrics continues. For instance, a counterfactual fairness definition has been developed [199] to capture the intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world—where the individual belongs to a different demographic group. Simulations can also be used to gain counterfactual information about how the data would have varied if a different data collection or decision-making policy had been in place [204]. As algorithmic discrimination can arise from the encoding of spurious correlations and noisy local dependencies into ML systems during training, there is currently great focus on causal tools [206] and how they can formally incorporate effects of hypothetical actions to solve a wide range of fairness modeling problems. Until causal methods are more widely available and adopted, minimizing the number of input variables, and ensuring that there is no strong correlation amongst them and a logical relationship to the prediction target, is a mitigation tactic for proxy discrimination and other AI risks.

[15] *See, e.g.,* U.S. v. Ga. Power Co., 474 F.2d 906 (5th Cir. 1973).

When deciding which fairness metric to adopt, it is important to recognize the impossibility of satisfying certain mathematical fairness constraints at once except in highly constrained special cases [207]. For example, there is an inherent incompatibility between two conditions: calibration and balancing the positive and negative classes. These conditions cannot be satisfied simultaneously unless under certain constraints [78]. While not all mathematical fairness desiderata can be achieved simultaneously, it is important to note that mitigated bias and good performance can be achieved simultaneously [208].

The plethora of fairness metric definitions illustrates that fairness cannot be reduced to a concise mathematical definition. Fairness is dynamic, social in nature, application and context specific, and not just an abstract or universal statistical problem. Therefore, it is important to adopt a socio-technical approach to fairness in order to have realistic fairness definitions for different contexts as well as task-specific datasets for machine learning model development and evaluation.

## 3.3 Who makes decisions and how do they make them? Human Factors in AI Bias

### 3.3.1 Human Factors Challenges

As ML algorithms have evolved in accuracy and precision, computational systems have moved from being used purely for decision support—or for explicit use by and under the control of a human operator—to automated decision making with limited input from humans. Computational decision support systems augment another, typically human, system in making decisions. Comparatively, for algorithmic decision systems there is less human involvement, with the AI system itself more in the "driver's seat," and able to produce outcomes with little human involvement to govern the impact. The growth and prevalence of algorithmic decision systems has helped to drive a decreased sense of trust in AI among the public [209]. This distrust is exacerbated by the reality that historical and social biases are baked-in to the data and assumptions used in the algorithmic models generating automated decisions. As a result, these algorithmic models have a higher probability of producing and amplifying unjust outcomes (e.g. for racial and ethnic minorities in areas such as criminal justice) [18–30, 210]. The systemic biases embedded in algorithmic models can also be exploited and used as a weapon at scale, causing catastrophic harm [211–214]. Organizations that deploy AI models and systems without assessing and managing these risks can not only harm their users but jeopardize their reputations.

**Deployment Context of Use**

AI systems are designed and developed to be used in specific real world settings, but are

often tested in idealized scenarios. Once deployed, the original intent, idea, or impact assessment can drift as the application is repurposed or used in unforeseen ways, and in settings or contexts for which it was not originally intended. Different deployment contexts means a new set of risks to be considered. Engaging with the broad set of stakeholder communities that may be impacted by the deployment of these technologies—before the decision is made to build the AI system—is an important consideration and strongly recommended. For more on context of use and what it encompasses from a human-centered design perspective, see subsequent Section 3.3.2.

One major purpose, and a significant benefit, of automated technology is that it can make sense of information more quickly and consistently than humans. AI systems are also often perceived as a way to make public interest decisions more fair, or to reduce (or eliminate) biased human decision making and bring about a more equitable society [27]. These perspectives have led to the deployment of automated and predictive modeling tools within trusted institutions and high-stakes settings such as hiring or criminal justice. In such settings, automated decisions that incorporate negative biases can perpetuate harms more quickly, extensively, and systematically than human and societal biases on their own.

**Human-in-the-loop**

Most algorithmic decision systems are socio-technical systems. They are inextricably tied to human social behavior, from the datasets used by ML processes and the decisions made by those who build them, to the interactions with the humans who provide the insight and oversight to make such systems actionable. The default assumption is that placing a human "in-the-loop" of such systems can ensure that adverse events do not occur. Current perceptions about the role and responsibility of the human-in-the-loop with AI are often implicit, and expectations about level of performance for these systems are often based on untested or outdated hypotheses. The bulk of academic literature available in this domain often relates to humans working with automated systems that pre-date the broad scale use of ML.

Some human-in-the-loop systems are deployed for use by subject matter experts. In this expert-driven scenario, professionals with expertise in a specific domain work in conjunction with an automated system towards a specific end goal—usually a consequential decision about another individual(s). Depending on the purpose of the system, the expert may interact with the ML model but is rarely part of the design or development of the system itself. These experts are not necessarily familiar with ML, data science, computer science, or other fields traditionally associated with AI design or development. For example, for AI systems that are deployed in the domain of medicine, the experts are the physicians and bring their expertise about medicine—not data science, data modeling and engineering, or other computational factors.

The perception that a human (expert or otherwise) can effectively and objectively oversee the use of algorithmic decision systems is a problematic assumption. More work needs to be done to understand the complex institutional and societal structures where these systems are developed and placed. Humans carry their own significant cognitive biases and HEURISTICS into the operation of AI systems and exactly how they can assist remains an understudied area. One challenge with human-in-the-loop scenarios is finding a configuration that enables a system to be used in a way that optimally leverages, instead of *replaces*, the subject matter expertise of the human. This is difficult since subject matter experts and AI developers often lack a common vernacular, which can contribute to miscommunication and misunderstood expectations and capabilities on both sides of the human-AI system.

**Expert-system** configurations are complex, even without the aid of a highly advanced AI. Experts and operators can often be placed into AI-based system settings without explicit declarations for governing authority over the specific task and outcome. With the promise of approaches that are more quantitative, subject matter experts may inadvertently activate the McNamara fallacy and leverage the AI system to take the pressure off of their often more subjective processes for the presumed objectivity of automation (this bias is often referred to as automation complacency). Expert users may also subconsciously find ways to leverage this perceived objectivity as cover, or even justification, for their implicit biases [215–217] and inadvertently make decisions that are inaccurate and harmful. Relatedly, AI developer communities may subconsciously presume that experts' methods have been validated to a greater degree than is the case. These kinds of implicit individual and group actions may create conditions that indirectly encourage the use of technology that is not quite ready for use, especially in high-stakes settings [3, 78, 218]. Researchers recommend that AI development teams work in tighter conjunction with subject matter experts and practitioner end users, who in turn, must "consider a deliberate and modest approach" when utilizing automated output [219].

Expert-driven ML and human-in-the-loop practices are not intended to serve as a form of oversight on AI systems and accompanying results. Experts bring their particular subject matter knowledge to the process, and are not necessarily trained to govern the use of an AI system they played no role in developing. But current legal and governance structures actively rely on humans—either expert or otherwise—to serve as a mechanism for protecting society from faulty, mistaken, and/or dangerous algorithmic decisions. The fundamental assumption of such structures is that a human overseer, simply by virtue of being human, will be able to provide adequate governance for systems.[16] The reality however is

[16]This is most frequently emphasized in governance frameworks that associate human-in-the-loop decisions as posing less risk, as opposed to fully automated decision making. See, for example, the role of general human intervention in minimizing risks for AI systems in the FDA's "Good Machine Learning Practice for Medical Device Development: Guiding Principles," https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-

that without significant procedural and cultural support, optimistic expectations about how humans are able to serve in this administrative capacity are not borne out in practice. The literature provides a thorough review of the flaws of human oversight policies [112].

**General public**
The challenge of interpretable systems is also a factor for consumer or citizen use of AI applications. It is presumed that trust can improve if the public is able to interrogate and engage with AI systems in a more transparent manner. In their article on public trust in AI, Knowles and Richards state "... members of the public do not need to trust individual AIs at all; what they need instead is the sanction of authority provided by suitably expert auditors that AI can be trusted" [220]. Developing such an authority requires standard practices, metrics, and norms from a socio-technical perspective. The NIST AI Risk Management Framework will help create standard practices, metrics and norms in consensus with the AI community.

Reliance on various downstream professionals to act as a governor on automated processes in complex societal systems is not a viable approach.

### 3.3.2 Human Factors Guidance

**Impact assessments**
The decision to deploy AI technology is a function of organizational incentives. AI is designed and developed within a set of organizational norms and policies. One recent proposed approach for ensuring that technology is developed in an ethical and responsible manner is the algorithmic impact assessment. Identifying and addressing potential biases is an important step in the assessment process. There is currently momentum for AI researchers to include statements about potential societal impacts [221] when submitting their work to journals or conferences. Similar to privacy impact assessments, which are relied upon by data protection and privacy frameworks to gauge and respond to data privacy risks, such impact assessments provide a high-level structure that enables organizations to frame the risks of each algorithm or deployment while also accounting for the specifics of each use case. Engaging in impact assessment can also serve as a forcing mechanism for

medical-device-development-guiding-principles; NHTSA's "Automated Driving Systems 2.0 Voluntary Guidance," https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf. In the military context, even more emphasis has been placed on human intervention, such as in "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence," Department of Defense Defense Innovation Board, https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF; see also Brig. Gen. (ret.) Jean Michel Verney et al., "Human-On-the-Loop," Joint Air & Space Power Conference 2021, https://www.japcc.org/human-on-the-loop/.

organizations to articulate any risks, and then to generate documentation of any mitigation activities in the event that any harms—and associated oversight—do arise[17][222–226]. A misstep with impact assessments is to only apply them once at the beginning of a long and iterative process in which goals and outcomes can change over time. To overcome the challenge of the point-in-time nature of impact assessments, impact assessments must be applied at some reasonable cadence when used with iterative and evolving AI systems. Another concern with impact assessments is that the technology groups, or others who will be assessed, may have undue influence on building or using the assessment.

**Multi-stakeholder engagement**

The practice of technology development is also complicated by the role of power and decision making within the organizational structure [227]. A consistent theme from the literature is the benefit of engaging a variety of stakeholders and maintaining diversity along social lines where bias is a concern (racial diversity, gender diversity, age diversity, diversity of physical ability) [228, 229]. These kinds of practices can lead to broadening perspectives, and in turn, more thorough evaluation of the societal impacts of technology-based applications. Using the demographic traits of organizational personnel to identify problematic aspects within development culture and practice is not sufficient and may not be fair. Identifying downstream impacts may take time and require the involvement of end-users, practitioners, subject matter experts, and interdisciplinary professionals from the law and social science. Expertise matters, and these stakeholders can bring their varied experiences to bear on the core challenge of identifying harmful outcomes and context shifts within the specific setting the AI system will be deployed.

Technology or datasets that seem non-problematic to one group may be deemed disastrous by others. The manner in which different user groups can game certain applications may also not be so obvious to the teams charged with bringing an AI-based technology to market. These kinds of impacts can sometimes be identified in early testing stages, but are usually very specific to the contextual end-use and will change over time. Acquiring these types of resources for risk and associated impacts does not necessarily require a huge allocation, but it does require deliberate planning and guidance. This is also a place where innovation in approaching bias could improve practice. These factors are part of changing norms and creating an organizational risk culture where teams improve capacity for considering the impact of the technology they design and develop, and communicating about these impacts more broadly.

**Diversity, Equity & Inclusion**

Without prioritizing diversity, equity, and inclusion in the teams involved in training and deploying AI systems it is difficult to move beyond a focus on system optimization or to address design considerations and risks beyond a narrow subset of users. Consider for example how character limits impact some languages and cultures more so than others; in

[17]H.R. 2231, 116th Cong. (2019), https://www.congress.gov/bill/116th-congress/house-bill/2231/text.

recognition of this effect, Twitter increased its character limit from 140 to 280 characters [230]. In another example, a recent exercise by the same social media company found that AI used to filter image content disfavored people with white hair and memes written in non-latin scripts [231, 232].

As recent research has shown that developers with similar demographic backgrounds make similar misjudgements, [72] ensuring that individuals involved in training, testing, and deploying the system have a diversity of experience, expertise and backgrounds is a critical risk mitigant that can help organizations manage the potential harms of AI. The human heuristics and biases that lead to examples such as these are implicit; as such, simply increasing awareness of bias does not ensure control over it. As previously described in Section 3.3, heuristics are adaptive mental shortcuts than can often be beneficial to reduce complexity in tasks of judgement and choice, yet also lead to cognitive biases.

The concepts and reasoning behind diversity, equity, and inclusion in the workplace are closely tied to the need for broad multi-stakeholder engagement during all aspects of the AI lifecycle. Numerous studies have touted the benefits of increased diversity, equity, and inclusion in the workplace [233–236]. Yet, the AI field noticeably lacks diversity [237]. To extend the benefits of diversity, equity, and inclusion to both the users and developers of AI systems, commentators and experts now recommend that bias mitigation efforts should be multifaceted, empowering a diverse group of individuals who reflect a range of backgrounds, perspectives and expertise, which in turn can help to broaden the views of AI system designers and engineers [238, 239]. In particular, diversity, equity and inclusion efforts can help organizations better understand: how the system is likely to impact a wide variety of users, how such users might interact with the system in practice, the potential harms and benefits of systems across users and groups, whether troubleshooting efforts—such as the recourse channels described below—are likely to be effective in practice, as well as how the system might impact broader populations beyond direct users of the system, among others.

**Practice Improvements**

By taking a lifecycle approach it is possible to identify junctures where well-developed guidance, assurance, and governance processes can assist business units and data and social scientists to collaboratively integrate processes to reduce bias without being cumbersome or blocking progress. Several technology companies are developing or utilizing guidance to improve organizational decision making and make the practice of AI development more responsible by implementing processes such as striving to identify potential bias impacts of algorithmic models. One approach is to enumerate institutional assumptions when developing algorithmic decision systems and map these assumptions to the expectations of the groups impacted by the technology–which requires deliberate multi-stakeholder and community engagement. "Cultural effective challenge" is a practice that seeks to create an environment where technology developers can actively challenge and question steps in modeling and engineering to help root out statistical biases and the biases inherent in human decision making [240]. Requiring AI practitioners to defend their techniques, within

a demographically and professionally diverse setting, can incentivize new ways of thinking, stimulate improved practices, and help create change in approaches by individuals and organizations [227].

**Human–AI configuration**

AI systems are often deliberately placed into high-risk settings to counteract the known subjectivity and bias of humans. Yet considerable questions remain about how to optimally configure humans and automation. An approach to human-in-the-loop that takes into consideration the broad set of socio-technical factors is necessary, especially in the context of AI bias. The list of relevant sub-topics span fields such as human factors, psychology, organizational behavior, and human-AI interaction, and building bridges between these and the technology communities is still necessary. NIST seeks to develop formal guidance about how to implement human-in-the-loop processes that do not amplify or perpetuate the many human, systemic and computational biases that can degrade outcomes in this complex setting. Identifying system configurations and necessary qualifications for their components that result in outcomes that are accurate and trustworthy will be a key focus.

**System and procedural transparency**

A consistent finding in the literature is that AI systems need to be more explainable and interpretable. The proliferation of tools such as datasheets and model cards are intended to fill that gap [241, 242]. Bias intersects with transparency in complex ways. Groups who invent and produce technology have specific intentions for its use and are unlikely to be aware of all the ways a given application will be used and repurposed once deployed. Transparency tools are especially helpful for addressing the problem of unintended use, but even when AI systems are used as intended there are significant individual differences in how humans interpret AI model output. This issue becomes particularly relevant when deploying systems for use by subject matter experts, who are less interested in *how* a system works and more concerned with *why* a system provided a given output. When system designers do not take these perceptual differences into consideration it can lead to misinterpretation of output, which is especially problematic in high-risk settings [243, 244]. Coordinated guidance is necessary to ensure that transparency tools are effectively supporting the professionals who use them and not indirectly contributing to processes that could amplify bias.

There are techniques to flag factors in datasets and modeling processes that can produce biased outcomes or cause noncompliance with legal requirements. The intent here is that flagging information for somebody along the AI lifecycle or the end user will serve as a system check. Yet, flagging such information for downstream users does not always result in a directly positive outcome, and can in fact create the opposite[181, 245]. Developing guidance in this area will require more information about the settings under which human biases may amplify harmful outcomes, and where humans can work optimally with and complement an AI-based system. These questions, like those related to AI system design, are notably dependent on setting (e.g., aircraft, cyber-physical systems, public safety and forensics, manufacturing), operator (e.g., expert, trained, naive), and task (e.g., recognition,

event detection, forecasting, reasoning).

**Keeping humans at the center of AI design**

Human-centered design (HCD) is an approach to the design and development of a system or technology that aims to improve the ability of users to effectively and efficiently use a product. HCD seeks to improve the user experience of an entire system, involving all aspects of a technology, from hardware design to software design. HCD is a methodology that has been successfully applied to a myriad of important domains, and NIST itself has authored several HCD handbooks tailored for particular domains, e.g.:, biometrics and public safety [246, 247].

HCD is an ongoing, iterative process in which project teams design, test, and continually refine a system, placing users at the core of the process. Humans and their needs drive the process, rather than having a techno-centric focus. HCD works as part of other development lifecycles, including waterfall, spiral and agile models. User-centered design, HCD, participatory design, co-design, and value-sensitive design all have key similarities; at the highest level, they seek to provide humans with designs that are ultimately beneficial to their lives. Furthermore, by placing humans at the center of such approaches, they naturally lend themselves to a deeper focus on larger societal considerations such as fairness, bias, values, and ethics. HCD works to create more usable products that meet the needs of its users. This, in turn, reduces the risk that the resulting system will under-deliver, pose risks to users, result in user harms, or fail.

The HCD process is illustrated in Fig. 6 below.

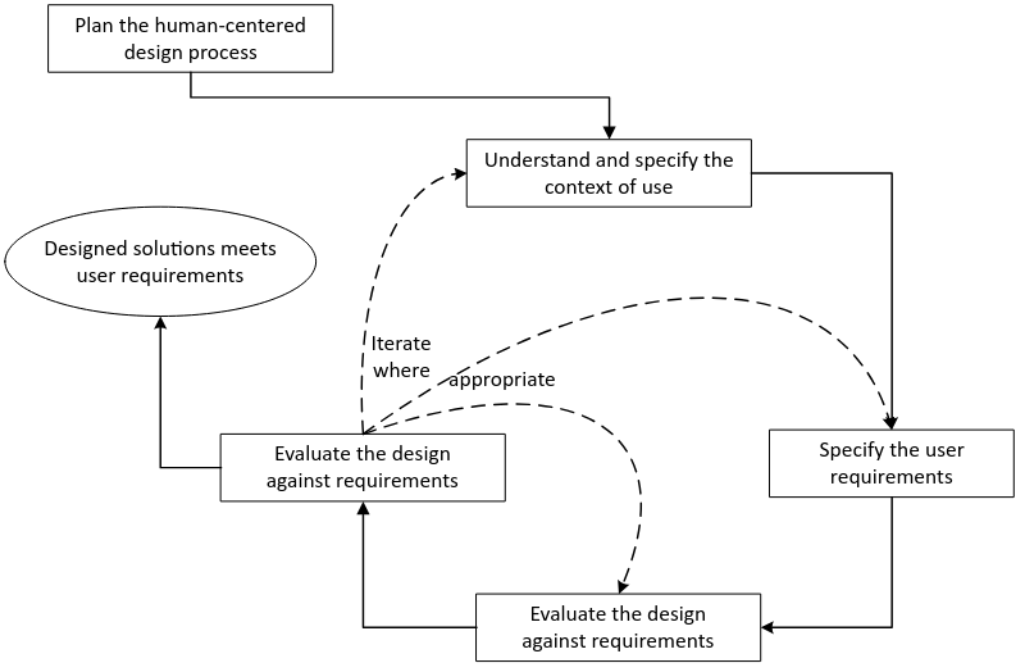


**Fig. 6.** Human-centered Design Process [ISO 9241-210:2019]

As defined in International Organization for Standardization (ISO) standard 9241-210:2019 [248], HCD involves:

- an explicit understanding of users, tasks and environments–the context of use;
- the involvement of users throughout design and development;
- a design driven and refined by human-centered evaluation;
- an iterative process whereby a prototype is designed, tested and modified;
- addressing the whole user experience;
- a design team including multidisciplinary skills and perspectives.

Based on the ISO standard, a HCD methodology for the development of AI systems could iteratively comprise the following, as shown in Fig. 7:

- Defining the Context of Use, including operational environment, user characteristics, tasks, and social environment;

- Determining the User & Organizational Requirements, including business requirements, user requirements, and technical requirements;
- Developing the Design Solution, including the system design, user interface, and training materials; and
- Conducting the Evaluation, including usability and conformance testing.

Although all components of HCD are critical, the context of use has key socio-technical considerations for AI systems. The socio-technical dynamics and conditions under which an AI system is used must be considered at the front end of any project to ensure that the design of the system will meet the needs of users, the objectives of the organization, and larger societal needs once the system is implemented in a real-world environment. A deep understanding of contextual factors is important throughout the AI lifecycle. Context of use does not simply involve the users' context of use, it involves a much broader view of context: the organizational environment in which the AI system is being developed (including existing systems and products); the operational environment in which the system will be used; and the larger societal environment in which the system will be implemented. For example, some intended users of AI systems may not have consistent or reliable access to fundamental internet technologies (a phenomenon widely described as the "digital divide" [249, 250]), leading to biases in how different communities access a system. Similarly, those with disabilities may experience difficulties interacting with AI systems. Crucially, such difficulties often cannot be mitigated by mathematical or software de-biasing approaches, and failure to address these important design issues may pose legal risks, for example in employment related activities affecting persons with disabilities.[18]

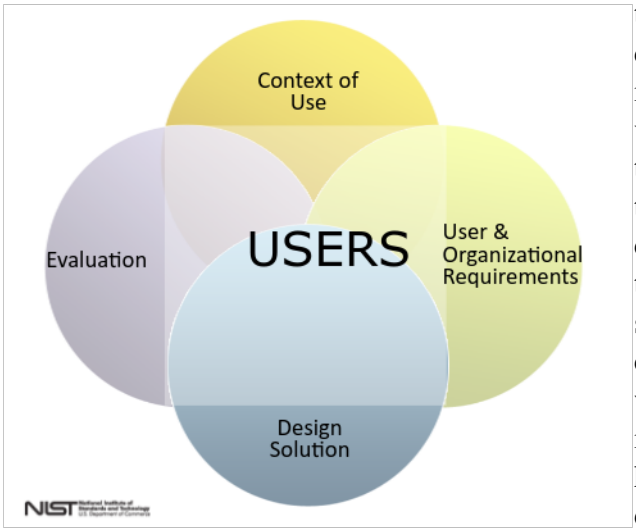


**Fig. 7.** Human-centered Design Process for AI Systems

[18]Congress has recognized that objects, systems, and processes often are not designed with individuals with disabilities in mind. By ensuring that these protections apply at the individual rather than group level, Congress further recognized that the means of placing an individual with disabilities on equal footing with others may require an individualized solution—one person with a disability may require a reasonable accommodation, and a different individual with a disability may require a different accommodation or no accommodation at all. Some disabilities are so heterogeneous that even two individuals with the same disability may need different accommodations. In the employment context, an algorithm may screen out a particular individual, and therefore may violate the Americans with Disabilities Act, regardless of whether broadly defined groups of individuals with disabilities tend to be assessed highly by a given algorithm.

A growing number of researchers have pointed out the benefits of socio-technical approaches. For example, Ferrer et al [251] note: "This challenge could be addressed through a socio-technical approach which can consider both the technical dimensions and the complex social contexts in which these systems are deployed. Building public confidence and greater democratic participation in AI systems requires ongoing development of not just explainable AI but of better Human-AI interaction methods and socio-technical platforms, tools and public engagement to increase critical public understanding and agency."

Research to integrate HCD with the standard design, development, evaluation, and deployment processes of today's AI systems is relatively recent. In their chapter on HCD of AI in the Handbook of Human Factors and Ergonomics, Margetis et al state that "A core concept of HCD is that of actively involving end-users and appropriate stakeholders in the process. In the context of AI, this means placing humans in the loop, not only through meaningful human control [252], but also through their active participation in the preparation, learning, and decision-making phases of AI [253]." Human-centered AI (HCAI) is an emerging area of scholarship that reconceptualizes HCD in the context of AI, providing human-centered AI design metaphors and suggested governance structures to develop reliable, safe, and trustworthy AI systems [254]. Schneiderman envisages HCAI as "bridg[ing] the gap between ethics and practice with specific recommendations for making successful technologies that augment, amplify, empower, and enhance humans rather than replace them. This shift in thinking could lead to a safer, more understandable, and more manageable future. An HCAI approach will reduce the prospects for out-of-control technologies, calm fears of robot-driven unemployment, and diminish the threats to privacy and security. A human-centered future will also support human values, respect human dignity, and raise appreciation for the human capacities that bring creative discoveries and inventions."

### 3.4 How do we manage and provide oversight? Governance and AI Bias

Governance processes impact nearly every aspect of managing AI bias. For that reason, it is essential to view governance as a holistic implementation tier, socio-technical in nature, and informing each phase of the bias management process. It is also important to note that governance does not simply focus on technical artifacts, such as AI systems alone, but also on organizational processes and cultural competencies that directly impact the individuals involved in training, deploying and monitoring such systems. While there are a number of components to effective governance for managing bias in AI systems, we focus here on organizational measures and culture.

#### 3.4.1 Governance Guidance

**Monitoring**

AI systems may perform differently than expected once deployed, which can lead to differential treatment of individuals from different groups. A key measure to control this risk is to deploy additional systems that monitor for potential bias issues, which can alert the proper personnel when potential problems are detected. Without such monitoring in place,

it can be difficult to know if deployed system performance in the real world matches up to the measurements conducted in a laboratory environment, or whether newly collected data match the distribution of the training data. A key consideration for the success of live monitoring for bias is the collection of data from the active user population, especially data related to user demographics such as age and gender, to enable calculation of assessment measures. These type of data can have a variety of privacy implications and may be subject to legal restrictions on what types of data can be collected and under what conditions.

**Recourse Channels**

Availability of feedback channels allow system end users to flag incorrect or potentially harmful results, and seek recourse for errors or harms. A number of legal frameworks prioritize the ability of users to appeal and override unfavorable decisions, and are applied in a subset of algorithmic systems deployed in areas like consumer finance. Because appeal and override recourse often requires a logical description of the questionable ML decision, these processes are tightly connected to AI system explainability and interpretability. Though not without criticism [255], adverse action notices for negative consumer credit decisions, as mandated by the Equal Credit Opportunity Act and the Fair Credit Reporting Act, are an example of an explanation and appeal process[19][256]. Additional appeal and override processes could include options for customers to interact with a human instead of an AI system or options to avoid similar AI-generated content in the future. Embedding such processes and technologies into AI systems allows users to appeal wrong decisions (or even suggestions) while also empowering technology development teams to remediate potential incidents at or near their inception point.

**Policies and Procedures**

In the context of AI systems, ensuring that written policies and procedures address key roles, responsibilities, and processes at all stages of the AI model lifecycle is critical to managing and detecting potential overall issues of AI system performance.[20] Policies and procedures can enable consistent development and testing practices, which in turn can help to ensure that results from AI systems are repeatable and that related risks are consistently mapped, measured and managed. Without such policies, the management of AI bias can easily become subjective and inconsistent across organizations, which can exacerbate risks over time rather than minimize them—if, for example, irreconcilably different metrics are used across systems. Policies may:

- define the key terms and concepts related to AI systems and the scope of their intended impact;
- address the use of sensitive or otherwise potentially risky data;

[19] *See* 15 U.S.C., § 1691(d).

[20] Bd. Governors Fed. Rsrv. Sys., Supervisory Guidance on Model Risk Management, SR Letter 11-7 (Apr. 4, 2011).

- detail standards for experimental design, data quality, and model training;
- outline how the risks of bias should be mapped and measured, and according to what standards;
- detail processes for model testing and validation;
- detail the process of review by legal or risk functions;
- set forth the periodicity and depth of ongoing auditing and review;
- outline requirements for change management; and
- detail any plans related to incident response for such systems, in the event that any significant risks do materialize during deployment.

**Documentation**

Clear documentation practices can help to systematically implement policies and procedures, standardizing how an organization's bias management processes are implemented and recorded at each stage. Standardized documentation can, in turn, help to ensure accountability, as described in further detail below. Model documents should contain interpretable descriptions of system mechanisms, enabling oversight personnel to make informed, risk-based decisions about the system's potential to perpetuate bias. Documentation also serves as a single repository for important information, supporting not only internal oversight of AI systems and related business processes, but also enhancing system maintenance, and serving as a valuable resource for any necessary corrective or debugging activities.[21]

Model documentation is especially important in the context of accountability. The use of documentation templates with specific requirements enables practitioners to walk through workflows as they are prescribed in written policies and procedures, or by other best practices. Omission of key documentation elements can indicate a lack of adherence to written policies and procedures on the part of system developers or testers. Some model documentation templates also include contact information for developers and stakeholders [241, 242]. The act of adding contact information to a document describing a work product can enable more efficient oversight and communications. This type of practice should also lead to greater concern and responsibility for the quality of the product, which in turn, can impact bias management efforts within an organization.

[21] Off. Comptroller Currency, Comptroller's Handbook: Model Risk Management (Aug. 2021), https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html.

**Accountability**

Accountability plays a critical role in governance efforts [257]. Governance without accountability is, in practice, unlikely to be effective. Ensuring that a specific team, and often, a specific individual – such as a Chief Model Risk Officer, as is now common in large consumer finance organizations – is responsible for bias management in AI systems is a fundamental accountability mechanism.[22] Ensuring individuals or teams bear responsibility for risks and associated harms provides a direct incentive for their mitigation. Put simply, when someone's boss is accountable for bias issues, they too are accountable for bias issues—and this phenomenon promulgates down to front-line practitioners. Accountability for AI bias cannot lie on the shoulders of a single individual, which is why accountability mandates should also be embedded within and across the various teams involved in the training and deployment of AI systems. Existing technical and procedural frameworks for accountability related to AI include general governance procedures, and application of system monitoring, data quality measures, computer security countermeasures, and nondiscrimination mechanisms, among others [258, 259].

Fundamentally, accountability requires a clear assessment of the role of the AI system itself. For example, decision-support systems, which may be claimed not to result in direct decision-making and therefore pose less risks, can easily become overly relied upon by users, or misused or abused. In these cases, the AI system would generate similar harms as if it were engaging in decision-making directly. Model or algorithmic audits [260] can be used to assess and document such crucial accountability considerations. There are several notions of audits commonly discussed in the responsible and trustworthy AI communities. Audit may refer to a traditional internal audit function employed to track issues of model risk, as in traditional model governance. Audit may refer to a structured and principled application of lessons learned in financial audit practices to AI systems [261]. Alternatively, audit may refer to some general documentation and transparency approach. Audits can be an effective accountability, bias, and general risk mitigation mechanism. Indeed, laws are being passed that demand bias audits of AI-based systems used in employment [262]. However, audits currently exist in a wide range of forms with varying levels of quality and consensus [263]. Audits will be addressed in future NIST documents related to the AI risk management framework.

**Culture & Practice**

For AI governance to be effective, it needs to be embedded throughout the culture of an organization. While organizational culture and practice can be defined in a variety of ways, the central theme of most such definitions emphasize beliefs, norms and values - or, in other words, the behavior an organization prioritizes in practice, even if such behavior is not codified or written down [264]. Risk management culture and practices can be a powerful technique for identifying biases across the AI lifecycle and from a socio-technical system perspective.

[22] Bd. Governors Fed. Rsrv. Sys., *supra* note 20.

**Effective challenge** The principal of effective challenge is a central component of model risk management frameworks. This practice is heavily relied on by the financial sector to mitigate algorithmic risk, and mandates that important model design and implementation decisions be questioned by experts with the authority and stature to make changes in design and implementation.[23] Fostering a culture of effective challenge encourages actively challenging and questioning steps in the development of AI systems, and can help to raise issues of AI bias before they materialize in deployed systems. An organizational culture that encourages serious questioning of AI system designs will be more likely to identify problems before they turn into harmful incidents. Relatedly, while individuals who are part of the development of AI systems may be knowledgeable about the potential harmful impacts of the technology they build, impact assessments should not be exclusively developed by these teams due to increased likelihood of confirmation bias and other incentives that may cause conflicts of interest.

**Three lines of defense** Because culture can be difficult to map or measure directly, one way to encourage this approach is to incentivize critical thinking and review at an organizational and procedural level. Model risk management frameworks, for example, are often systematically implemented through the so-called "three lines of defense," which creates separate teams that are held accountable for different aspects of the model lifecycle. Typically, the first line of defense focuses on model development, the second on risk management, and the third on auditing.[24] While a traditional three-lines approach may be impractical for smaller organizations, ensuring that a culture of effective challenge is encouraged and sustained can help organizations to anticipate, and therefore to effectively mitigate, risks of bias before they materialize.

**Risk Mitigation, Risk Tiering & Incentive Structures**

Some applications of AI are high-risk.[25] A central cultural component of effective risk management for AI bias lies in a clear acknowledgment that risk mitigation, rather than risk avoidance, is often the most effective factor in managing such risks.[26] Developing a risk mitigation mindset, meaning a clear acceptance that incidents can and will occur, and emphasizing practical detection and mitigation once they do, can help ensure that any risks of bias are quickly mitigated in practice. This acknowledgement enables a clear triaging of risks which can enable organizations to focus finite resources on the risks of bias that are most material, and therefore most likely to cause real-world harm. An additional component of effective organizational culture includes aligning pay and promotion incentives across teams to AI risk mitigation efforts, such that participants in the risk mitigation

[23] *Id.*

[24] Off. Superintendent Fin. Inst. Canada, Enterprise-Wide Model Risk Management for Deposit-Taking Institutions, E-23 (Sept. 2017).

[25] Eur. Comm'n, Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (proposed Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[26] Bd. Governors Fed. Rsrv. Sys., *supra* note 20

mechanisms—like the three lines of defense—are truly motivated to use sound development approaches, test rigorously and audit thoroughly.[27]

**Information Sharing**

As described in a NIST special publication [265], sharing cyber threat information helps organizations improve both their own security postures, and those of other organizations. Identifying internal mechanisms for teams to share information about bias incidents or other harmful impacts from AI helps to elevate the importance of AI risks and provides information for teams to avoid past failed designs. Some initial efforts are already underway [266]. As teams begin to create norms for tracking such incidents, it can potentially transform AI practices and the organizational culture. Improving awareness of how bias presents in deployed AI and its related impacts can enhance knowledge and capabilities, and prevent incidents. Fostering a culture of information sharing can also serve as a new area for community engagement.

## 4. Conclusions

This document has provided a broad overview of the complex challenge of addressing and managing risks associated with AI bias. It is clear that developing detailed technical guidance to address this challenging area will take time and input from diverse stakeholders, within and beyond those groups who design, develop, and deploy AI applications, and including members of communities that may be impacted by the deployment of AI systems.

Since AI is neither built nor deployed in a vacuum, we approach AI as a socio-technical system, acknowledging that AI systems and associated bias extend beyond the computational level. Bias can be introduced purposefully or inadvertently, or it can emerge as the AI system is used, impacting society at large through perpetuating and amplifying biased and discriminatory outcomes. Adopting a socio-technical perspective brings new requirements, many of which are contextual in nature, to the processes that comprise the AI lifecycle. It is important to gain understanding in how computational and statistical factors interact with systemic and human biases.

NIST has provided an initial socio-technical framing for AI bias in this document, including key context and terminology, highlights of the main challenges, and foundational directions for future guidance. This information is classified and discussed through the document according to three key areas:

1. dataset availability, representativeness, and suitability in socio-technical contexts;
2. TEVV considerations for measurement and metrics to support testing and evaluation;
3. human factors, including societal and historic biases within individuals and organizations, participatory approaches such as human-centered design, and human–in–the–loop practices.

[27] *Id.*

Identifying the key requirements for improving our knowledge in this area is a necessary first step. To ensure broad input, engagement, and consensus, NIST will carry out supporting standards development activities such as workshops and public comment periods for draft documents.

NIST intends to develop further consensus socio-technical guidance in collaboration with the research community and a broad set of other stakeholders, including those who are directly impacted by AI bias. The intent is for this guidance to be of specific assistance for organizations who commission, design, develop, deploy, use, or evaluate AI for a variety of use cases. By providing these entities with clear, explicit, and technically valid guidance NIST intends to improve the state of practice for AI bias and assure system trustworthiness.

## 5. Glossary

**activity bias** A type of selection bias that occurs when systems/platforms get their training data from their most active users, rather than those less active (or inactive) [131]. 8

**aleatoric uncertainty** Aleatoric uncertainty, also known as statistical uncertainty, refers to unknowns that differ each time we run the same experiment. It refers to the variability in the outcome of an experiment which is due to inherently random effects. For example, in machine learning context, the data-generating process may have a stochastic component that cannot be reduced by any additional source of information. Consequently, even the best model trained on this data will not be able to provide a definite answer. 9, 20, 21

**amplification bias** Arises when the distribution over prediction outputs is skewed in comparison to the prior distribution of the prediction target [267]. 8

**anchoring bias** A cognitive bias, the influence of a particular reference point or anchor on people's decisions. Often more fully referred to as anchoring-and-adjustment, or anchoring-and-adjusting: after an anchor is set, people adjust insufficiently from that anchor point to arrive at a final answer. Decision makers are biased towards an initially presented value [79]. 8, 9

**annotator reporting bias** When users rely on automation as a heuristic replacement for their own information seeking and processing [268]. A form of individual bias but often discussed as a group bias, or the larger effects on natural language processing models. 8

**automation complacency** When humans over-rely on automated systems or have their skills attenuated by such over-reliance (e.g., spelling and autocorrect or spellcheckers). 8

**availability heuristic** Also referred to as availability bias. A mental shortcut whereby people tend to overweight what comes easily or quickly to mind, meaning that what is easier to recall—e.g., more "available"—receives greater emphasis in judgement and decision-making. 8

**behavioral bias** Systematic distortions in user behavior across platforms or contexts, or across users represented in different datasets [144, 269]. 8

**cognitive bias** A broad term referring generally to a systematic pattern of deviation from rational judgement and decision-making. A large variety of cognitive biases have been identified over many decades of research in judgement and decision-making, some of which are adaptive mental shortcuts known as heuristics. 8

**concept drift** Use of a system outside the planned domain of application, and a common cause of performance gaps between laboratory settings and the real world. 8

**confirmation bias** also called confirmatory bias, a cognitive bias where people tend to prefer information that aligns with, or confirms, their existing beliefs. People can exhibit confirmation bias in the search for, interpretation of, and recall of information. In the famous Wason selection task experiments, participants repeatedly showed a preference for confirmation over falsification. They were tasked with identifying an underlying rule that applied to number triples they were shown, and they overwhelmingly tested triples that confirmed rather than falsified their hypothesized rule [270]. 8, 9, 27

**construct validity** A form of validation that seeks to answer whether a test measures what it intends to measure. [271]. 15

**consumer bias** Arises when an algorithm or platform provides users with a new venue within which to express their biases, and may occur from either side, or party, in a digital interaction [272]. 8

**content production bias** Arises from structural, lexical, semantic, and syntactic differences in the contents generated by users [144]. 8

**data dredging** A statistical bias in which testing huge numbers of hypotheses of a dataset may appear to yield statistical significance even when the results are statistically nonsignificant. 8, 27

**data generation bias** Arises from the addition of synthetic or redundant data samples to a dataset [273]. 8

**deployment bias** Arises when systems are used as decision aids for humans, since the human intermediary may act on predictions in ways that are typically not modeled in the system [90]. However, it is still individuals using the deployed system. 8, 26

**detection bias** Systematic differences between groups in how outcomes are determined and may cause an over- or underestimation of the size of the effect [274]. 8

**Dunning–Kruger effect** A cognitive bias, the tendency of people with low ability in a given area or task to overestimate their self-assessed ability. Typically measured by comparing self-assessment with objective performance, often called subjective ability and objective ability, respectively [275]. 8, 26

**ecological fallacy** Occurs when an inference is made about an individual based on their membership within a group. 8, 23

**emergent bias** Use of a system outside the planned domain of application, and a common cause of performance gaps between laboratory settings and the real world. 8

**epistemic uncertainty** An epistemic uncertainty, also known as systematic uncertainty, refers to deficiencies by a lack of knowledge or information. This may be because the methodology on which a model is built neglects certain effects or because particular data have been deliberately hidden. 9, 20–22

**error propagation** Arises when applications that are built with machine learning are used to generate inputs for other machine learning algorithms. If the output is biased in any way, this bias may be inherited by systems using the output as input to learn other models [82]. 8

**evaluation bias** Arises when the testing or external benchmark populations do not equally represent the various parts of the user population or from the use of performance metrics that are not appropriate for the way in which the model will be used [90]. 8

**exclusion bias** When specific groups of user populations are excluded from testing and subsequent analyses [276]. 8

**feedback loop bias** Effects that may occur when an algorithm learns from user behavior and feeds that behavior back into the model [272]. 8

**funding bias** Arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study [85], but it can also be the individual researcher. 8

**governance** a framework of policies, rules, and processes for ensuring direction, management and accountability. ii

**groupthink** A psychological phenomenon that occurs when people in a group tend to make non-optimal decisions based on their desire to conform to the group, or fear of dissenting with the group. In groupthink, individuals often refrain from expressing their personal disagreement with the group, hesitating to voice opinions that do not align with the group. 8

**heuristics** in the context of human decision making, often referred to as "mental shortcuts," a term that encompasses many methods that may be less than fully rational or optimal, yet are often sufficient for an approximate solution. Although heuristics can reduce cognitive load and aid people when making decisions, such heuristics also result in systematic errors and cognitive biases [79]. 34

**historical bias** referring to the long-standing biases encoded in society over time. Related to, but distinct from, biases in historical description, or the interpretation, analysis, and explanation of history. A common example of historical bias is the tendency to view the larger world from a Western or European view. 8

**human reporting bias** When users rely on automation as a heuristic replacement for their own information seeking and processing [268]. 8

**implicit bias** An unconscious belief, attitude, feeling, association, or stereotype that can affect the way in which humans process information, make decisions, and take actions. 8

**inherited bias** Arises when applications that are built with machine learning are used to generate inputs for other machine learning algorithms. If the output is biased in any way, this bias may be inherited by systems using the output as input to learn other models [82]. 8

**institutional bias** In contrast to biases exhibited at the level of individual persons, institutional bias refers to a tendency exhibited at the level of entire institutions, where practices or norms result in the favoring or disadvantaging of certain social groups. Common examples include institutional racism and institutional sexism [91]. 8

**interpretation bias** A form of information processing bias that can occur when users interpret algorithmic outputs according to their internalized biases and views [272]. 8

**language model** A computational model that has been trained using statistical methods to find patterns in written and/or spoken language, in order to predict or classify words, text, or speech. 21

**linking bias** Arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users [144]. 8

**loss of situational awareness bias** When automation leads to humans being unaware of their situation such that, when control of a system is given back to them in a situation where humans and machines cooperate, they are unprepared to assume their duties. This can be a loss of awareness over what automation is and isn't taking care of. 8

**McNamara fallacy** The belief that quantitative information is more valuable than other information. 12

**measurement bias** Arises when features and labels are proxies for desired quantities, potentially leaving out important factors or introducing group or input-dependent noise that leads to differential performance [90]. 8

**mode confusion bias** When modal interfaces confuse human operators, who misunderstand which mode the system is using, taking actions which are correct for a different mode but incorrect for their current situation. This is the cause of many deadly accidents, but also a source of confusion in everyday life. 8

**model** A conceptual, mathematical, or physical representation of phenomenon observed in a system of ideas, events, or processes. In computationally-based models used in AI, phenomenon are often abstracted for mathematical representation, which means that characteristics that can not be represented mathematically may not be captured in the model. i, v

**model selection bias** The bias introduced while using the data to select a single seemingly "best" model from a large set of models employing many predictor variables. Model selection bias also occurs when an explanatory variable has a weak relationship with the response variable [277]. 8

**popularity bias** A form of selection bias that occurs when items that are more popular are more exposed and less popular items are under-represented [130]. 8

**population bias** Systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population [144]. 8

**presentation bias** Biases arising from how information is presented on the Web, via a user interface, due to rating or ranking of output, or through users' own self-selected, biased interaction [131]. 8

**proxy** A variable that can stand in for another, usually not directly observable or measurable, variable. 20

**ranking bias** A form of anchoring bias. The idea that top-ranked results are the most relevant and important and will result in more clicks than other results [131, 278]. 8

**Rashomon effect or principle** Refers to differences in perspective, memory and recall, interpretation, and reporting on the same event from multiple persons or witnesses. 8

**representation bias** Arises due to non-random sampling of subgroups, causing trends estimated for one population to not be generalizable to data collected from a new population [85]. 8

**selective adherence** Decision-makers' inclination to selectively adopt algorithmic advice when it matches their pre-existing beliefs and stereotypes [215]. 8

**Simpson's Paradox** A statistical phenomenon where the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables. For example, the statistical association or correlation that has been detected between two variables for an entire population disappears or reverses when the population is divided into subgroups. 8, 17

**societal bias** often referred to as social bias. Can be positive or negative, and take a number of different forms, but is typically characterized as being for or against groups or individuals based on social identities, demographic factors, or immutable physical characteristics. Societal or social biases are often stereotypes. Common examples of societal or social biases are based on concepts like race, ethnicity, gender, sexual orientation, socioeconomic status, education, and more. Societal bias is often recognized and discussed in the context of NLP (Natural Language Processing) models. 8

**socio-technical** A term used to describe how humans interact with technology within the broader societal context. ii

**streetlight effect** A bias whereby people tend to search only where it is easiest to look [279]. 8

**sunk cost fallacy** A human tendency where people opt to continue with an endeavor or behavior due to previously spent or invested resources, such as money, time, and effort, regardless of whether costs outweigh benefits. For example, in AI, the sunk cost fallacy could lead development teams and organizations to feel that because they have already invested so much time and money into a particular AI application, they must pursue it to market rather than deciding to end the effort, even in the face of significant technical debt and/or ethical debt. 8

**survivorship bias** tendency for people to focus on the items, observations, or people that "survive" or make it past a selection process, while overlooking those that did not. 8

**technochauvinism** The belief that technology is always the solution [35]. 12

**temporal bias** Bias that arises from differences in populations and behaviors over time [144, 280]. 8

**uncertainty bias** Arises when predictive algorithms favor groups that are better represented in the training data, since there will be less uncertainty associated with those predictions [281]. 8

**user interaction bias** Arises when a user imposes their own self-selected biases and behavior during interaction with data, output, results, etc [131]. 8

## References


[1] NIST, "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools," National Institute of Standards and Technology, Tech. Rep., 2019. [Online]. Available: https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

[2] I. Ajunwa, S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "Hiring by Algorithm: Predicting and Preventing Disparate Impact," *undefined*, 2016. [Online]. Available: /paper/Hiring-by-Algorithm%3A-Predicting-and-Preventing-Ajunwa-Friedler/bd31ad5e998629998f35db9a10d858b36e603248

[3] S. Barocas, A. Biega, B. Fish, J. Niklas, and L. Stark, "When not to design, build, or deploy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, p. 695. [Online]. Available: https://doi.org/10.1145/3351095.3375691

[4] M. Bogen, "All the Ways Hiring Algorithms Can Introduce Bias," *Harvard Business Review*, May 2019, section: Hiring. [Online]. Available: https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

[5] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[6] E. Harlen and O. Schnuck, "Objective or Biased," 2021. [Online]. Available: https://web.br.de/interaktiv/ki-bewerbung/en

[7] J. Sanchez-Monedero, L. Dencik, and L. Edwards, "What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems," *arXiv:1910.06144 [cs]*, Jan. 2020, arXiv: 1910.06144. [Online]. Available: http://arxiv.org/abs/1910.06144

[8] M. Evans and A. W. Mathews, "New York Regulator Probes UnitedHealth Algorithm for Racial Bias," *Wall Street Journal*, Oct. 2019. [Online]. Available: https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601

[9] H. Fry, *Hello world: being human in the age of algorithms*. WW Norton & Company, 2018.

[10] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Intern Med*, vol. 178, no. 11, p. 1544, Nov. 2018. [Online]. Available: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2018.3763

[11] E. Guo and K. Hao, "This is the Stanford vaccine algorithm that left out frontline doctors," 2020. [Online]. Available: https://www.technologyreview.com/2020/12/21/1015303/stanford-vaccine-algorithm/

[12] H. Ledford, "Millions of black people affected by racial bias in health-care algorithms," *Nature*, vol. 574, no. 7780, pp. 608–609, Oct. 2019,

number: 7780 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/d41586-019-03228-6

[13] T. M. Maddox, J. S. Rumsfeld, and P. R. O. Payne, "Questions for Artificial Intelligence in Health Care," *JAMA*, vol. 321, no. 1, p. 31, Jan. 2019. [Online]. Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2018.18932

[14] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aax2342

[15] T. Simonite, "How an Algorithm Blocked Kidney Transplants to Black Patients | WIRED," *Wired*, 2020. [Online]. Available: https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/

[16] M. Singh and K. N. Ramamurthy, "Understanding racial bias in health using the Medical Expenditure Panel Survey data," *arXiv:1911.01509 [cs, stat]*, Nov. 2019, arXiv: 1911.01509. [Online]. Available: http://arxiv.org/abs/1911.01509

[17] T. M. Cruz, "Perils of data-driven equity: Safety-net care and big data's elusive grasp on health inequality," *Big Data & Society*, vol. 7, no. 1, p. 205395172092809, Jan. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2053951720928097

[18] J. Angwin, J. Larson, S. Mattu, L. Kirchner, and ProPublica, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*, 2016.

[19] L. Dormehl, "Algorithms Are Great and All, But They Can Also Ruin Lives," *Wired*, 2014. [Online]. Available: https://www.wired.com/2014/11/algorithms-great-can-also-ruin-lives/

[20] S. Goel, R. Shroff, J. L. Skeem, and C. Slobogin, "The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment," *SSRN Journal*, 2018. [Online]. Available: https://www.ssrn.com/abstract=3306723

[21] S. Brayne, "Enter the Dragnet," 2020. [Online]. Available: https://logicmag.io/commons/enter-the-dragnet/

[22] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017. [Online]. Available: http://www.liebertpub.com/doi/10.1089/big.2016.0047

[23] EPIC, "Algorithms in the Criminal Justice System Risk Assessment Tools," Electronic Privacy Information Center (EPIC), Tech. Rep., 2020. [Online]. Available: https://epic.org/algorithmic-transparency/crim-justice/

[24] K. Hill, "Flawed Facial Recognition Leads To Arrest and Jail for New Jersey Man," 2020. [Online]. Available: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

[25] J. E. Johndrow and K. Lum, "An algorithm for removing sensitive information: Application to race-independent recidivism prediction," *Ann. Appl. Stat.*, vol. 13,

no. 1, pp. 189–220, Mar. 2019. [Online]. Available: https://projecteuclid.org/euclid.aoas/1554861646

[26] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan, "Classifying Socially Sensitive Data Without Discrimination: An Analysis of a Crime Suspect Dataset," in *2012 IEEE 12th International Conference on Data Mining Workshops*. Brussels, Belgium: IEEE, Dec. 2012, pp. 370–377. [Online]. Available: http://ieeexplore.ieee.org/document/6406464/

[27] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, vol. 133, no. 1, pp. 237–293, 2018.

[28] A. Liptak, "Sent to Prison by a Software Program's Secret Algorithms," *The New York Times*, May 2017. [Online]. Available: https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html

[29] R. Wexler, "When a Computer Program Keeps You in Jail," *The New York Times*, p. 2, 2017. [Online]. Available: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

[30] "State v. Loomis," 2016.

[31] M. Aitken, E. Toreini, P. Carmichael, K. Coopamootoo, K. Elliott, and A. van Moorsel, "Establishing a social licence for Financial Technology: Reflections on the role of the private sector in pursuing ethical data practices," *Big Data & Society*, vol. 7, no. 1, p. 205395172090889, Jan. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2053951720908892

[32] J. P. Bajorek, "Voice Recognition Still Has Significant Race and Gender Biases," *Harvard Business Review*, 2019, section: Technology. [Online]. Available: https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

[33] E. Bary, "How artificial intelligence could replace credit scores and reshape how we get loans," 2018. [Online]. Available: https://www.marketwatch.com/story/ai-based-credit-scores-will-soon-give-one-billion-people-access-to-banking-services-2018-10-09

[34] R. Benjamin, *Race after technology: Abolitionist tools for the new jim code*. John Wiley & Sons, 2019.

[35] M. Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

[36] J. Boulamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of Machine Learning Research*, 2018, pp. 77–91.

[37] C. Criado-Perez, *Invisible Women: Data Bias in a World Designed for Men*. Abrams Press, 2019.

[38] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," *arXiv:1104.3913 [cs]*, Nov. 2011, arXiv: 1104.3913. [Online]. Available: http://arxiv.org/abs/1104.3913

[39] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish*

*the poor*. St. Martin's Press, 2018.
[40] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *arXiv:1610.02413 [cs]*, Oct. 2016, arXiv: 1610.02413. [Online]. Available: http://arxiv.org/abs/1610.02413
[41] S. Noble, *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.
[42] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2017.
[43] A. Pandey and A. Caliskan, "Iterative Effect-Size Bias in Ridehailing: Measuring Social Bias in Dynamic Pricing of 100 Million Rides," *arXiv:2006.04599 [cs]*, Jun. 2020, arXiv: 2006.04599. [Online]. Available: http://arxiv.org/abs/2006.04599
[44] J. Redden, "The Harm That Data Do," *Scientific American*, 2018. [Online]. Available: https://www.scientificamerican.com/article/the-harm-that-data-do/
[45] M. Specia, "Siri and Alexa Reinforce Gender Bias, U.N. Finds," *The New York Times*, May 2019. [Online]. Available: https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html
[46] D. M. West, "Brookings survey finds worries over AI impact on jobs and personal privacy, concern U.S. will fall behind China," Brookings, Tech. Rep., 2018.
[47] S. Furman and J. Haney, "Is My Home Smart or Just Connected?" in *International Conference on Human-Computer Interaction*. Cham: Springer, 2020, pp. 273–287.
[48] A. Kerr, M. Barry, and J. D. Kelleher, "Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance," *Big Data & Society*, vol. 7, no. 1, p. 205395172091593, Jan. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2053951720915939
[49] E. Fast and E. Horvitz, "Long-Term Trends in the Public Perception of Artificial Intelligence," *Thirty-First AAAI Conference on Artificial Intelligence*, p. 7, 2017.
[50] A. Smith and M. Anderson, "Automation in Everyday Life," Pew Research Center, Tech. Rep., 2017.
[51] W. H. Ware, "Records, Computers and the Rights of Citizens," RAND Corporation, Santa Monica, CA, Tech. Rep., 1973. [Online]. Available: https://www.rand.org/pubs/papers/P5077.html.Alsoavailableinprintform.
[52] T. Feathers, "Major Universities Are Using Race as a "High Impact Predictor" of Student Success – The Markup," 2021, section: News. [Online]. Available: https://themarkup.org/news/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success
[53] L. Kirchner and M. Goldstein, "Access Denied: Faulty Automated Background Checks Freeze Out Renters – The Markup," 2020, section: Locked Out. [Online]. Available: https://themarkup.org/locked-out/2020/05/28/access-denied-faulty-automated-background-checks-freeze-out-renters
[54] I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," *Cardozo L. Rev.*, vol. 41, p. 1671, 2020. [Online]. Available: https://heinonline.org/HOL/Page?handle=hein.journals/cdozo41&id=1711&div=&collection=

[55] M. Bogen and A. Rieke, "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias," 2019. [Online]. Available: https://www.upturn.org/reports/2018/hiring-algorithms

[56] H. Schellmann, "Auditors are testing hiring algorithms for bias, but big questions remain," 2021. [Online]. Available: https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/

[57] R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-Lending Discrimination in the FinTech Era," National Bureau of Economic Research, Tech. Rep. w25943, Jun. 2019. [Online]. Available: https://www.nber.org/papers/w25943

[58] M. Henry-Nickie, "How artificial intelligence affects financial consumers," 2019. [Online]. Available: https://www.brookings.edu/research/how-artificial-intelligence-affects-financial-consumers/

[59] M. Weber, M. Yurochkin, S. Botros, and V. Markov, "Black Loans Matter: Distributionally Robust Fairness for Fighting Subgroup Discrimination," *arXiv:2012.01193 [cs]*, Nov. 2020, arXiv: 2012.01193. [Online]. Available: http://arxiv.org/abs/2012.01193

[60] H. Suresh and J. V. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," *arXiv:1901.10002 [cs, stat]*, Jun. 2021, arXiv: 1901.10002. [Online]. Available: http://arxiv.org/abs/1901.10002

[61] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019.

[62] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Online edition, 2021, 4th US edition, http://aima.cs.berkeley.edu/index.html.

[63] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.

[64] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *AI & SOCIETY*, vol. 35, 12 2020.

[65] K. De Vries, "Identity, Profiling Algorithms and a World of Ambient Intelligence," *Ethics and Information Technology*, vol. 12, pp. 71–85, 03 2010.

[66] R. Richardson, J. M. Schultz, and K. Crawford, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, And Justice," *New York University Law Review*, vol. 94, p. 42, 2018.

[67] D. Elliott, R. G. Lowitz, and W. C. NFP, "What Is the Cost of Poor Credit?" *Washington, DC: Urban Institute*, 2018.

[68] W. Haven, "Bias Isn't the Only Problem with Credit Scores - and, No, AI Can't Help," MIT Technology Review, June 2021, https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/.

[69] L. Sarkesian and S. Singh, "HUD's New Rule Paves the Way for Rampant Algorithmic Discrimination in Housing Decisions," New America, Oct. 2020, https://www.newamerica.org/oti/blog/huds-new-rule-paves-the-way-for-rampant-algorithmic-discrimination-in-housing-decisions/.

[70] OECD, "Glossary of statistical terms," OECD Online Resource, July 2007, https://stats.oecd.org/glossary/detail.asp?ID=3605.

[71] ISO, "Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability," ISO, Tech. Rep. ISO 3534-1:2006, 2006. [Online]. Available: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/04/01/40145.html

[72] B. Cowgill, F. Dell'Acqua, S. Deng, D. Hsu, N. Verma, and A. Chaintreau, "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics," *arXiv:2012.02394 [cs, econ, q-fin]*, Dec. 2020, arXiv: 2012.02394. [Online]. Available: http://arxiv.org/abs/2012.02394

[73] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016, publisher: California Law Review, Inc. [Online]. Available: https://www.jstor.org/stable/24758720

[74] S. Costanza-Chock, "Design Justice, A.I., and Escape from the Matrix of Domination," *Journal of Design and Science*, Jul. 2018. [Online]. Available: https://jods.mitpress.mit.edu/pub/costanza-chock

[75] M. Elish, S. Barocas, A. Plasek, and K. Ferryman, "The social & economic implications of artificial intelligence technologies in the near-term," AI Now, New York, Tech. Rep., 2016. [Online]. Available: https://ainowinstitute.org/AI_Now_2016_Primers.pdf

[76] A. Jacobs and H. Wallach, "Measurement and Fairness," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, Mar. 2021, arXiv: 1912.05511. [Online]. Available: http://arxiv.org/abs/1912.05511

[77] S. Passi and S. Barocas, "Problem Formulation and Fairness," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 39–48, Jan. 2019, arXiv: 1901.02547. [Online]. Available: http://arxiv.org/abs/1901.02547

[78] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. Atlanta, GA, USA: ACM Press, 2019, pp. 59–68. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3287560.3287598

[79] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.jstor.org/stable/1738360

[80] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, Apr. 2017. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aal4230

[81] D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia: International Joint Conferences on Artificial

Intelligence Organization, Aug. 2017, pp. 4691–4697. [Online]. Available: https://www.ijcai.org/proceedings/2017/654

[82] T. Hellström, V. Dignum, and S. Bensch, "Bias in Machine Learning – What is it Good for?" *arXiv:2004.00686 [cs]*, Sep. 2020, arXiv: 2004.00686. [Online]. Available: http://arxiv.org/abs/2004.00686

[83] ISO/IEC, "ISO/IEC 2382:2015, Information technology — Vocabulary," International Organization for Standardization, Geneva, Switzerland, Tech. Rep., 2015. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en

[84] ——, "ISO/IEC 20546:2019, Information technology — Big data — Overview and vocabulary," International Organization for Standardization, Geneva, Switzerland, Tech. Rep., 2019. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:20546:ed-1:v1:en

[85] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *arXiv:1908.09635 [cs]*, Sep. 2019, arXiv: 1908.09635. [Online]. Available: http://arxiv.org/abs/1908.09635

[86] M. Mitchell, *Artificial Intelligence: A Guide for Thinking Human*. Farrar, Straus, and Giroux, 2019.

[87] S. Mitchell and J. Shadlen, "Mirror mirror: Reflections on quantitative fairness. Shira Mitchell: Statistician," Dec 2020. [Online]. Available: https://shiraamitchell.github.io/fairness/

[88] D. K. Mulligan, J. A. Kroll, N. Kohli, and R. Y. Wong, "This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–36, Nov. 2019, arXiv: 1909.11869. [Online]. Available: http://arxiv.org/abs/1909.11869

[89] Organisation for Economic Co-operation and Development, "Recommendation of the Council on Artificial Inteliigence," 2019. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[90] H. Suresh and J. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," *arXiv:1901.10002 [cs, stat]*, Feb. 2020, arXiv: 1901.10002. [Online]. Available: http://arxiv.org/abs/1901.10002

[91] D. Chandler and R. Munday, *A Dictionary of Media and Communication*. Oxford University Press, Jan. 2011, publication Title: A Dictionary of Media and Communication. [Online]. Available: https://www.oxfordreference.com/view/10.1093/acref/9780199568758.001.0001/acref-9780199568758

[92] M. Ngan, P. J. Grother, and M. Ngan, *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.

[93] D. I. Perrett, K. J. Lee, I. Penton-Voak, D. Rowland, S. Yoshikawa, D. M. Burt, S. Henzi, D. L. Castles, and S. Akamatsu, "Effects of sexual dimorphism on facial attractiveness," *Nature*, vol. 394, no. 6696, pp. 884–887, 1998.

[94] I. M. Scott, A. P. Clark, S. C. Josephson, A. H. Boyette, I. C. Cuthill, R. L. Fried, M. A. Gibson, B. S. Hewlett, M. Jamieson, W. Jankowiak *et al.*, "Human prefer-

ences for sexually dimorphic faces may be evolutionarily novel," *Proceedings of the National Academy of Sciences*, vol. 111, no. 40, pp. 14 388–14 393, 2014.

[95] K. Kleisner, P. Tureček, S. C. Roberts, J. Havlíček, J. V. Valentova, R. M. Akoko, J. D. Leongómez, S. Apostol, M. A. Varella, and S. A. Saribay, "How and why patterns of sexual dimorphism in human faces vary across the world," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[96] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, 2018. [Online]. Available: https://doi.org/10.1145/3274357

[97] Organization of Scientific Area Committees for Forensic Science, "OSAC Preferred Terms," 2021. [Online]. Available: https://www.nist.gov/system/files/documents/2021/04/28/OSAC%20Preferred%20Terms_April%202021.pdf

[98] D. Kahneman, S. P. Slovic, P. Slovic, and A. Tversky, *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[99] A. I. Al-Alawi, M. Naureen, E. I. AlAlawi, and A. A. N. Al-Hadad, "The Role of Artificial Intelligence in Recruitment Process Decision-Making," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2021, pp. 197–203.

[100] A. Rieke, U. Janardan, M. Hsu, and N. Duarte, "Essential Work: Analyzing the Hiring Technologies of Large Hourly Employers," Upturn, July 2021, https://www.upturn.org/reports/2021/essential-work/.

[101] L. X. Z. Brown and M. Richardson, "Algorithm-driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination?" Ctr. for Democracy & Tech., Dec. 2020, https://cdt.org/insights/report-algorithm-driven-hiring-tools-innovative-recruitment-or-expedited-disability-discrimination.

[102] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, and D. Sculley, "Underspecification Presents Challenges for Credibility in Modern Machine Learning," *arXiv:2011.03395 [cs, stat]*, Nov. 2020, arXiv: 2011.03395. [Online]. Available: http://arxiv.org/abs/2011.03395

[103] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Kohd, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar,

F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," *arXiv:2108.07258 [cs]*, Aug. 2021, arXiv: 2108.07258. [Online]. Available: http://arxiv.org/abs/2108.07258

[104] D. Schiff, A. Ayesh, L. Musikanski, and J. C. Havens, "IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence," *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2020. [Online]. Available: http://dx.doi.org/10.1109/SMC42975.2020.9283454

[105] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, "The Values Encoded in Machine Learning Research," *arXiv:2106.15590 [cs]*, Jun. 2021, arXiv: 2106.15590. [Online]. Available: http://arxiv.org/abs/2106.15590

[106] N. Schmidt and B. Stephens, "An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination," *arXiv preprint arXiv:1911.05755*, 2019.

[107] C. Barabas, C. Doyle, J. Rubinovitz, and K. Dinakar, "Studying Up: Reorienting the Study of Algorithmic Fairness Around Issues of Power," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 167–176.

[108] B. Fish and L. Stark, "Reflexive Design for Fairness and Other Human Values in Formal Models," *arXiv:2010.05084 [cs]*, Oct. 2020, arXiv: 2010.05084. [Online]. Available: http://arxiv.org/abs/2010.05084

[109] K. Robertson, C. Khoo, and Y. Song, "To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada," Citizen Lab & Int'l Hum. Rts. Prog., U. Toronto, Sept. 2020, https://citizenlab.ca/wp-content/uploads/2020/09/To-Surveil-and-Predict.pdf.

[110] S. C. Slota, K. R. Fleischmann, S. Greenberg, N. Verma, B. Cummings, L. Li, and C. Shenefiel, "Many hands make many fingers to point: challenges in creating accountable AI," *AI & Soc*, Nov. 2021. [Online]. Available: https://link.springer.com/10.1007/s00146-021-01302-0

[111] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions," *Annu. Rev. Stat. Appl.*, vol. 8, no. 1, pp. 141–163, Mar. 2021, arXiv: 1811.07867. [Online]. Available: http://arxiv.org/abs/1811.07867

[112] B. Green, "The Flaws of Policies Requiring Human Oversight of Government Algorithms," *SSRN Journal*, 2021. [Online]. Available: https://www.ssrn.com/abstract=3921216

[113] B. Green and A. Kak, "The false comfort of human oversight as an antidote to A.I. harm." [Online]. Available: https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html
[114] M. Boyarskaya, A. Olteanu, and K. Crawford, "Overcoming Failures of Imagination in AI Infused System Development and Deployment," *arXiv:2011.13416 [cs]*, Dec. 2020, arXiv: 2011.13416. [Online]. Available: http://arxiv.org/abs/2011.13416
[115] D. Boyd and K. Crawford, "CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, Jun. 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878
[116] C. D'Ignazio and L. Klein, *Data Feminism*. MIT Press, 2020. [Online]. Available: https://data-feminism.mitpress.mit.edu/
[117] A. Jacobs, S. L. Blodgett, S. Barocas, H. Daumé, and H. Wallach, "The meaning and measurement of bias: lessons from natural language processing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, p. 706. [Online]. Available: https://doi.org/10.1145/3351095.3375671
[118] E. Moss and J. Metcalf, "High Tech, High Risk: Tech Ethics Lessons for the COVID-19 Pandemic Response," *Patterns*, vol. 1, no. 7, p. 100102, Oct. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666389920301367
[119] A. L. Washington and R. Kuo, "Whose side are ethics codes on?: power, responsibility and the social good," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, Jan. 2020, pp. 230–240. [Online]. Available: https://dl.acm.org/doi/10.1145/3351095.3372844
[120] E. Morozov, *To save everything, click here: The folly of technological solutionism*. Penn State University Press, 2013.
[121] B. Aguera y Arcas, M. Mitchell, and A. Todorov, "Physiognomy's new clothes," May 2017. [Online]. Available: https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a
[122] J. A. Kroll, "Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 758–771, Mar. 2021, arXiv: 2101.09385. [Online]. Available: http://arxiv.org/abs/2101.09385
[123] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," 2017.
[124] R. Tromble, "Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age," *Social Media + Society*, vol. 7, no. 1, p. 2056305121988929, Jan. 2021, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/2056305121988929
[125] A. Cobham, *The Uncounted*. John Wiley & Sons, 2020.
[126] D. Stone, *Counting: How We Use Numbers to Decide What Matters*. Liveright Publishing, 2020.

[127] B. Plank, "What to do about non-standard (or non-canonical) language in NLP," *arXiv:1608.07836 [cs]*, Aug. 2016, arXiv: 1608.07836. [Online]. Available: http://arxiv.org/abs/1608.07836

[128] Y. Tan and L. E. Celis, "Assessing Social and Intersectional Biases in Contextualized Word Representations," *arXiv:1911.01485 [cs, stat]*, Nov. 2019, arXiv: 1911.01485. [Online]. Available: http://arxiv.org/abs/1911.01485

[129] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *arXiv:2012.05345 [cs]*, Dec. 2020, arXiv: 2012.05345. [Online]. Available: http://arxiv.org/abs/2012.05345

[130] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher, "The Unfairness of Popularity Bias in Recommendation," *arXiv:1907.13286 [cs]*, Sep. 2019, arXiv: 1907.13286. [Online]. Available: http://arxiv.org/abs/1907.13286

[131] R. Baeza-Yates, "Bias on the web," *Commun. ACM*, vol. 61, no. 6, pp. 54–61, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3209581

[132] A. Lambrecht and C. E. Tucker, "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *SSRN Journal*, 2016. [Online]. Available: http://www.ssrn.com/abstract=2852260

[133] M. Miceli, J. Posada, and T. Yang, "Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?" *arXiv:2109.08131 [cs]*, Sep. 2021, arXiv: 2109.08131. [Online]. Available: http://arxiv.org/abs/2109.08131

[134] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951. [Online]. Available: http://www.jstor.org/stable/2984065

[135] F. P. Calmon, D. Wei, K. N. Ramamurthy, and K. R. Varshney, "Optimized data pre-processing for discrimination prevention," 2017.

[136] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012. [Online]. Available: https://doi.org/10.1007/s10115-011-0463-8

[137] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," 2015. [Online]. Available: https://arxiv.org/abs/1412.3756

[138] K. Peng, A. Mathur, and A. Narayanan, "Mitigating dataset harms requires stewardship: Lessons from 1000 papers," *arXiv:2108.02922 [cs]*, Aug. 2021, arXiv: 2108.02922. [Online]. Available: http://arxiv.org/abs/2108.02922

[139] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," *arXiv:1810.01943 [cs]*, Oct. 2018, arXiv: 1810.01943. [Online]. Available: http://arxiv.org/abs/1810.01943

[140] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets desta-

bilize machine learning benchmarks," 2021.
[141] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Quantifying and Reducing Stereotypes in Word Embeddings," *arXiv:1606.06121*, vol. 1, p. 5, 2016.
[142] P. Parasurama and J. Sedoc, "Gendered Language in Resumes and its Implications for Algorithmic Bias in Hiring," *arXiv:2112.08910 [cs]*, Dec. 2021, arXiv: 2112.08910. [Online]. Available: http://arxiv.org/abs/2112.08910
[143] A. L. Hoffmann, "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication & Society*, vol. 22, no. 7, pp. 900–915, 2019.
[144] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Front. Big Data*, vol. 2, p. 13, Jul. 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdata.2019.00013/full
[145] T. N. Bond and K. Lang, "The sad truth about happiness scales," *Journal of Political Economy*, vol. 127, no. 4, pp. 1629–1640, 2019.
[146] D. Kahneman, A. M. Rosenfield, L. Gandhi, and T. Blaser, "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making," *Harvard Business Review*, Oct. 2016, section: Decision making and problem solving. [Online]. Available: https://hbr.org/2016/10/noise
[147] M. M. Malik, "A Hierarchy of Limitations in Machine Learning," *arXiv:2002.05193 [cs, econ, math, stat]*, Feb. 2020, arXiv: 2002.05193. [Online]. Available: http://arxiv.org/abs/2002.05193
[148] G. Friedman and T. McCarthy, "Employment Law Red Flags in the Use of Artificial Intelligence in Hiring," 2019. [Online]. Available: https://www.americanbar.org/groups/business_law/publications/blt/2020/10/ai-in-hiring/
[149] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, p. 457–506, Mar 2021. [Online]. Available: http://dx.doi.org/10.1007/s10994-021-05946-3
[150] Y. Nesterov, *Introductory lectures on convex optimization*, 2004th ed., ser. Applied Optimization. New York, NY: Springer, Dec. 2003.
[151] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for Optimization*. London, England: MIT Press, Mar. 2019.
[152] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.
[153] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2021.
[154] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21.

New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623. [Online]. Available: https://doi.org/10.1145/3442188.3445922

[155] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, "It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia," in *Proceedings of the international AAAI conference on web and social media*, vol. 9, no. 1, 2015, pp. 454–463.

[156] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial intelligence and statistics*. PMLR, 2015, pp. 192–204.

[157] IEEE, "IEEE Standard for Floating-Point Arithmetic," *IEEE Std 754-2008*, pp. 1–70, 2008.

[158] K. Novak, *Numerical Methods for Scientific Computing: The Definitive Manual for Math Geeks*, 2nd ed. Equal Share Press, 2022.

[159] M. J. Wolf, K. W. Miller, and F. S. Grodzinsky, "Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications," *The ORBIT Journal*, vol. 1, no. 2, pp. 1–12, 2017.

[160] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving, "Scaling language models: Methods, analysis & insights from training gopher," 2022.

[161] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," 2021.

[162] A. F. Ansari, M. L. Ang, and H. Soh, "Refining deep generative models via discriminator gradient flow," 2021.

[163] J. Z. Forde, A. F. Cooper, K. Kwegyir-Aggrey, C. De Sa, and M. Littman, "Model Selection's Disparate Impact in Real-World Deep Learning Applications," *arXiv:2104.00606 [cs]*, Apr. 2021, arXiv: 2104.00606. [Online]. Available: http://arxiv.org/abs/2104.00606

[164] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of "Bias" in NLP," *arXiv:2005.14050 [cs]*, May 2020, arXiv: 2005.14050. [Online]. Available: http://arxiv.org/abs/2005.14050

[165] G. Neff and P. Nagy, "Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay," *International Journal of*

*Communication*, vol. 10, no. 0, p. 17, Oct. 2016, number: 0. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/6277

[166] D. Hovy and S. Prabhumoye, "Five sources of bias in natural language processing," *Language and Linguistics Compass*, vol. 15, no. 8, p. e12432, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12432. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432

[167] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut Learning in Deep Neural Networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[168] M. Hashemi and M. Hall, "RETRACTED ARTICLE: Criminal Tendency Detection from Facial Images and the Gender Bias Effect," *Journal of Big Data*, vol. 7, no. 1, pp. 1–16, 2020.

[169] BBC, "Facial recognition to 'predict criminals' sparks row over AI bias," BBC Online News, June 2020, https://www.bbc.com/news/technology-53165286.

[170] S. Levin, "New AI can guess whether you're gay or straight from a photograph," The Guardian, Sept. 2017, https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph.

[171] J. D. West and C. T. Bergstrom, "Misinformation In and About Science," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, 2021.

[172] A. E. Miller, "Searching for gaydar: Blind spots in the study of sexual orientation perception," *Psychology & Sexuality*, vol. 9, no. 3, pp. 188–203, 2018.

[173] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[174] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews," in *Proceedings of the 2021 International Conference on Multimodal Interaction*. Montréal QC Canada: ACM, Oct. 2021, pp. 268–277. [Online]. Available: https://dl.acm.org/doi/10.1145/3462244.3479897

[175] A. Narayanan, "How to recognize AI snake oil," CITP (Princeton U.), Feb. 2022, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

[176] T. H. Davenport and J. G. Harris, *Competing on analytics*. Boston, MA: Harvard Business Review Press, Feb. 2007.

[177] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein, "Algorithms as discrimination detectors," *Proc Natl Acad Sci USA*, vol. 117, no. 48, pp. 30 096–30 100, Dec. 2020. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1912790117

[178] M. H. Jarrahi, G. Newlands, M. K. Lee, C. T. Wolf, E. Kinder, and W. Sutherland, "Algorithmic management in a work context," *Big Data & Society*, vol. 8, no. 2, p. 20539517211020332, 2021.

[179] S. Brayne, *Predict and surveil*. New York, NY: Oxford University Press, Jan. 2021.

[180] B. Cowgill, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening," *Columbia Business School, Columbia University*, p. 35, 2020.
[181] C. Thompson, "Who's homeless enough for housing? in san francisco an algorithm decides," Nov 2021. [Online]. Available: https://www.codastory.com/authoritarian-tech/san-francisco-homeless-algorithm/
[182] Office of Inspector General, "Advisory concerning the Chicago police department's predictive risk models," City of Chicago, Tech. Rep., Jan. 2020, https://igchicago.org/wp-content/uploads/2020/01/OIG-Advisory-Concerning-CPDs-Predictive-Risk-Models-.pdf.
[183] D. Hunter and N. Evans, "Facebook emotional contagion experiment controversy," *Research Ethics*, vol. 12, no. 1, pp. 2–3, 2016.
[184] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology*, vol. 114, p. 246–257, 2018.
[185] M. Roberts, D. Driggs, M. Thorpe, J. D. Gilbey, M. Yeung, S. Ursprung, A. I. Avilés-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, and C.-B. Schönlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *Nat. Mach. Intell.*, vol. 3, pp. 199–217, 2021.
[186] L. Wynants, B. Calster, G. Collins, R. Riley, G. Heinze, E. Schuit, M. Bonten, D. Dahly, J. Damen, T. Debray, V. Jong, M. Vos, P. Dhiman, M. Haller, M. Harhay, L. Henckaerts, P. Heus, M. Kammer, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, G. Martin, D. McLernon, C. Navarro, J. Reitsma, J. Sergeant, C. Shi, N. Skoetz, L. Smits, K. Snell, M. Sperrin, R. Spijker, E. Steyerberg, T. Takada, I. Tzoulaki, S. Kuijk, B. Bussel, I. Horst, F. Royen, J. Verbakel, C. Wallisch, J. Wilkinson, R. Wolff, L. Hooft, K. Moons, and M. Smeden, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, p. m1328, Apr. 2020, publisher: British Medical Journal Publishing Group Section: Research. [Online]. Available: https://www.bmj.com/content/369/bmj.m1328
[187] M. Hutson, *Has artificial intelligence become alchemy?* American Association for the Advancement of Science, 2018.
[188] R. Dijkgraaf, Quanta Magazine, Oct 2021. [Online]. Available: https://www.quantamagazine.org/science-has-entered-a-new-era-of-alchemy-good-20211020/
[189] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *bmj*, vol. 369, 2020.
[190] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association

for Computational Linguistics, Jul. 2019, pp. 1668–1678. [Online]. Available: https://aclanthology.org/P19-1163

[191] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab, "Bias in Data-driven AI Systems - An Introductory Survey," *arXiv:2001.09762v1 [cs.CY]*, p. 19, 2020.

[192] "Leveraging responsible AI to counteract bias in health care," Aug. 2021. [Online]. Available: https://www.statnews.com/2021/08/06/leverage-responsible-ai-counteract-bias-health-care/

[193] A. Słowik and L. Bottou, "Algorithmic Bias and Data Bias: Understanding the Relation between Distributionally Robust Optimization and Data Curation," *arXiv:2106.09467 [cs, stat]*, Jun. 2021, arXiv: 2106.09467. [Online]. Available: http://arxiv.org/abs/2106.09467

[194] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

[195] CFPB, "Using publicly available information to proxy for unidentified race and ethnicity," Consumer Financial Protection Bureau, 2014, https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

[196] N. Gill, P. Hall, K. Montgomery, and N. Schmidt, "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing," *Information*, vol. 11, no. 3, p. 137, 2020.

[197] S. Venkatasubramanian, C. Scheidegger, S. Friedler, and A. Clauset, *Fairness in Networks: Social Capital, Information Access, and Interventions*. New York, NY, USA: Association for Computing Machinery, 2021, p. 4078–4079. [Online]. Available: https://doi.org/10.1145/3447548.3470821

[198] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making," *Commun. ACM*, vol. 64, no. 4, p. 136–143, mar 2021. [Online]. Available: https://doi.org/10.1145/3433949

[199] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," *arXiv:1703.06856 [cs, stat]*, Mar. 2018, arXiv: 1703.06856. [Online]. Available: http://arxiv.org/abs/1703.06856

[200] L. Wang, "Race as proxy: Situational racism and self-fulfilling stereotypes," *DePaul Law Review*, vol. 53, p. 1013, 2004.

[201] A. Agan and S. Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment*," *The Quarterly Journal of Economics*, vol. 133, no. 1, pp. 191–235, 08 2017. [Online]. Available: https://doi.org/10.1093/qje/qjx028

[202] K. Fiscella and A. M. Fremont, "Use of geocoding and surname analysis to estimate race and ethnicity," *Health services research*, vol. 41, pp. 1482–500, 2006.

[203] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in Reinforcement Learning," *arXiv:1611.03071 [cs]*, Aug. 2017, arXiv: 1611.03071.

[Online]. Available: http://arxiv.org/abs/1611.03071
[204] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 525–534. [Online]. Available: https://doi.org/10.1145/3351095.3372878
[205] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.
[206] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," 2018.
[207] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," 2016.
[208] K. T. Rodolfa, H. Lamba, and R. Ghani, "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 896–904, 2021.
[209] R. Richardson, "Defining and demystifying automated decision systems," March 2021, forthcoming. [Online]. Available: https://ssrn.com/abstract=3D3811708
[210] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions," Montreal QC, Canada, Jan. 2018. [Online]. Available: https://osf.io/9wqxr
[211] C. M. Scaparrotti, *Joint Publication 3-13 Information Operations*. Citeseer, 2012.
[212] G. Raman, B. AlShebli, M. Waniek, T. Rahwan, and J. C.-H. Peng, "How weaponizing disinformation can bring down a city's power grid," *PLOS ONE*, vol. 15, pp. 1–14, 08 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0236517
[213] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, p. 96–104, jun 2016. [Online]. Available: https://doi.org/10.1145/2818717
[214] W. Phillips, "The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online," Data & Society, May 2018, https://datasociety.net/wp-content/uploads/2018/05/FULLREPORT_Oxygen_of_Amplification_DS.pdf.
[215] S. Alon-Barkat and M. Busuioc, "Decision-makers Processing of AI Algorithmic Advice: Automation Bias versus Selective Adherence," *arXiv:2103.02381 [cs]*, Mar. 2021, arXiv: 2103.02381. [Online]. Available: http://arxiv.org/abs/2103.02381
[216] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: people erro-

neously avoid algorithms after seeing them err," *J Exp Psychol Gen*, vol. 144, no. 1, pp. 114–126, Feb. 2015.

[217] ——, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science*, vol. 64, no. 3, pp. 1155–1170, Mar. 2018. [Online]. Available: http://pubsonline.informs.org/doi/10.1287/mnsc.2016.2643

[218] M. Veale, M. Van Kleek, and R. Binns, "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 - CHI '18*. Montreal QC, Canada: ACM Press, 2018, pp. 1–14. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3173574.3174014

[219] S. Picard, M. Watkins, M. Rempal, and A. Kerodal, "Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness," Center for Court Innovation, Tech. Rep., 2020. [Online]. Available: https://www.courtinnovation.org/publications/beyond-algorithm

[220] B. Knowles and J. T. Richards, "The Sanction of Authority: Promoting Public Trust in AI," *arXiv:2102.04221 [cs]*, Jan. 2021, arXiv: 2102.04221. [Online]. Available: http://arxiv.org/abs/2102.04221

[221] C. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, and A. Dafoe, "Institutionalizing ethics in AI through broader impact requirements," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 104–110, Feb. 2021, number: 2 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-021-00298-y

[222] E. Moss, E. Watkins, R. Singh, M. Elish, and J. Metcalf, "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." [Online]. Available: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

[223] G. Can., "Algorithmic Impact Assessment Tool," Gov't Can. Online Resource, Apr. 2021, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[224] M. Kop, "AI Impact Assessment & Code of Conduct," Futurium, May 2019, https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/ai-impact-assessment-code-conduct.

[225] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability," AI Now, Apr. 2018, https://ainowinstitute.org/aiareport2018.pdf.

[226] A. D. Selbst, "An Institutional View Of Algorithmic Impact Assessments," *Harvard Journal of Law & Technology*, vol. 35, no. 1, 2021.

[227] E. Moss and J. Metcalf, "Ethics Owners," Sep. 2020, publisher: Data & Society Research Institute. [Online]. Available: https://datasociety.net/library/ethics-owners/

[228] K. Crawford, "Artificial Intelligence's White Guy Problem," *The New York Times*,

p. 2, Jun. 2016. [Online]. Available: https://nyti.ms/28YaKg7
[229] D. Rock and H. Grant, "Why Diverse Teams Are Smarter," Nov. 2016, https://hbr.org/2016/11/why-diverse-teams-are-smarter.
[230] A. Rosen and I. Ihara, "Giving you more characters to express yourself," Twitter Blog, Sept. 2017, https://blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.
[231] W. Knight, "Twitter's Photo-Cropping Algorithm Favors Young, Thin Females," Wired, Aug. 2021, https://www.wired.com/story/twitters-photo-cropping-algorithm-favors-young-thin-females.
[232] K. Yee and I. F. Peradejordi, "Sharing learnings from the first algorithmic bias bounty challenge," Twitter Engineering Blog, Sept. 2021, https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge.
[233] C. Herring, "Does diversity pay?: Race, gender, and the business case for diversity," *American Sociological Review*, vol. 74, pp. 208 – 224, 2009.
[234] N. Ellemers and F. Rink, "Diversity in work groups," *Current opinion in psychology*, vol. 11, pp. 49–53, 2016.
[235] K. Talke, S. Salomo, and A. Kock, "Top management team diversity and strategic innovation orientation: The relationship and consequences for innovativeness and performance," *Journal of Product Innovation Management*, vol. 28, pp. 819–832, 2011.
[236] R. Lorenzo and M. Reeves, "How and Where Diversity Drives Financial Performance," Harvard Bus. Rev., Jan. 2018, https://hbr.org/2018/01/how-and-where-diversity-drives-financial-performance.
[237] S. M. West, M. Whittaker, and K. Crawford, "Discriminating Systems: Gender, Race, and Power in AI," AI Now Institute, Tech. Rep., 2019. [Online]. Available: https://ainowinstitute.org/discriminatingsystems.pdf
[238] D. Walsh, "How can human-centered ai fight bias in machines and people?" MIT Sloan Mgmt. Rev., Feb. 2021, https://mitsloan.mit.edu/ideas-made-to-matter/how-can-human-centered-ai-fight-bias-machines-and-people.
[239] M. Li, "To Build Less-Biased AI, Hire a More Diverse Team," Harvard Bus. Rev., Oct. 2020, https://hbr.org/2020/10/to-build-less-biased-ai-hire-a-more-diverse-team.
[240] P. Hall, N. Gill, and B. Cox, *Responsible machine learning: Actionable strategies for mitigating risks and driving adoption*. Sebastopol, CA: O'Reilly Media Inc., 2020.
[241] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. Atlanta, GA, USA: ACM Press, 2019, pp. 220–229. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3287560.3287596
[242] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach,

H. Daumee III, and K. Crawford, "Datasheets for Datasets," *arXiv:1803.09010v6 [cs.DB]*, 2020.
[243] D. Broniatowski, "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence," NIST, Tech. Rep., 2021.
[244] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi, "Do as AI say: susceptibility in deployment of clinical decision-aids," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–8, Feb. 2021. [Online]. Available: https://www.nature.com/articles/s41746-021-00385-9
[245] J. Zerilli, A. Knott, J. MacLaurin, and C. Gavaghan, "Algorithmic decision-making and the control problem," *Minds and Machines*, vol. 29, pp. 555 – 578, 2019.
[246] NIST, "Usability and Biometrics: Ensuring Successful Biometric Systems," NIST Online Resource, June 2008, https://www.nist.gov/system/files/usability_and_biometrics_final2.pdf.
[247] M. Theofanos *et al.*, "Usability Handbook for Public Safety Communications: Ensuring Successful Systems for First Responders," NIST (Handbook 161), May 2017, https://nvlpubs.nist.gov/nistpubs/hb/2017/NIST.HB.161.pdf.
[248] ISO, "Ergonomics of human-system interaction — Part 210: Human-centered design for interactive systems," ISO 9241-210:2019 (2nd ed.), July 2019, https://www.iso.org/standard/77520.html.
[249] E. A. Vogels, "Some digital divides persist between rural, urban and suburban America," Pew Research Center, Aug. 2021, https://www.pewresearch.org/fact-tank/2021/08/19/some-digital-divides-persist-between-rural-urban-and-suburban-america/.
[250] ——, "Digital divide persists even as americans with lower incomes make gains in tech adoption," Pew Research Center, June 2021, https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/.
[251] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and Discrimination in AI: a cross-disciplinary perspective," *IEEE Technol. Soc. Mag.*, vol. 40, no. 2, pp. 72–80, Jun. 2021, arXiv: 2008.07309. [Online]. Available: http://arxiv.org/abs/2008.07309
[252] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *AI Magazine*, vol. 36, no. 4, pp. 105–114, Dec. 2015. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2577
[253] G. Margetis, S. Ntoa, M. Antona, and C. Stephanidis, *HUMAN-CENTERED DESIGN OF ARTIFICIAL INTELLIGENCE*. John Wiley & Sons, Ltd, 2021, ch. 42, pp. 1085–1106. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119636113.ch42
[254] B. Shneiderman, *Human-Centered AI*. London, England: Oxford University Press, Jan. 2022.
[255] S. Ejaz, "A Broken System: How the Credit Reporting System Fails Consumers and What To Do About It," Consumer Reports, June 2021, https://advocacy.consumerreports.org/wp-content/uploads/2021/06/A-Broken-System-How-the-Credit-Repor

ting-System-Fails-Consumers-and-What-to-Do-About-It.pdf.
[256] S. Ammermann, "Adverse Action Notice Requirements Under the ECOA and the FCRA," Consumer Compliance Outlook, 2nd Q. 2013, https://consumercomplianceoutlook.org/2013/second-quarter/adverse-action-notice-requirements-under-ecoa-fcra.
[257] A. Smith, "Using Artificial Intelligence and Algorithms," FTC, Apr. 2020, https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms.
[258] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable Algorithms," *University of Pennsylvania Law Review*, vol. 165, no. 633, p. 74, 2017.
[259] GAO, "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities," GAO@100 (GAO-21-519SP), June 2021, https://www.gao.gov/assets/gao-21-519sp.pdf.
[260] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 33–44. [Online]. Available: https://doi.org/10.1145/3351095.3372873
[261] R. Carrier and S. Brown, "Taxonomy: AI Audit, Assurance, and Assessment," ForHumanity, Feb. 2021, https://forhumanity.center/web/wp-content/uploads/2021/09/ForHumanity.center_Taxonomy_AI_Audit_Assurance_Assessment.pdf.
[262] E. Mulvaney, "NYC Targets Artificial Intelligence Bias in Hiring Under New Law," Bloomberg Law, 2021, https://news.bloomberglaw.com/daily-labor-report/nyc-targets-artificial-intelligence-bias-in-hiring-under-new-law.
[263] R. N. Landers and T. S. Behrend, "Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes ai predictive models," 2022.
[264] D. Sull, S. Turconi, and C. Sull, "When It Comes to Culture, Does Your Company Walk the Talk?" MIT Sloan Mgmt. Rev., July 2020, https://sloanreview.mit.edu/article/when-it-comes-to-culture-does-your-company-walk-the-talk.
[265] C. Johnson, M. Badger, D. Waltermire, J. Snyder, and C. Skorupka, "Guide to cyber threat information sharing," National Institute of Standards and Technology, NIST Special Publication 800-150, Nov 2016. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-150
[266] S. McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database," *arXiv:2011.08512 [cs]*, Nov. 2020, arXiv: 2011.08512. [Online]. Available: http://arxiv.org/abs/2011.08512
[267] K. Leino, E. Black, M. Fredrikson, S. Sen, and A. Datta, "Feature-Wise Bias Amplification," *arXiv:1812.08999 [cs, stat]*, Oct. 2019, arXiv: 1812.08999. [Online]. Available: http://arxiv.org/abs/1812.08999
[268] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, "Seeing through the

Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2930–2939. [Online]. Available: http://ieeexplore.ieee.org/document/7780689/

[269] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, ""blissfully happy" or "ready to fight": Varying interpretations of emoji," in *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*. AAAI press, Jan. 2016, pp. 259–268. [Online]. Available: https://experts.umn.edu/en/publications/blissfully-happy-or-ready-to-fight-varying-interpretations-of-emo

[270] P. C. Wason, "Reasoning about a rule," *Quarterly Journal of Experimental Psychology*, vol. 20, no. 3, pp. 273–281, 1968. [Online]. Available: https://doi.org/10.1080/14640746808400161

[271] L. J. Cronbach and P. E. Meehl, "Construct validity in psychological tests." *Psychological bulletin*, vol. 52 4, pp. 281–302, 1955.

[272] S. Silva and M. Kenney, "Algorithms, platforms, and ethnic bias," *Commun. ACM*, vol. 62, no. 11, pp. 37–39, Oct. 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3318157

[273] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data," *arXiv:1811.11479 [cs, stat]*, Nov. 2018, arXiv: 1811.11479. [Online]. Available: http://arxiv.org/abs/1811.11479

[274] Centre for Evidence-Based Medicine, "Catalogue of Bias," Mar. 2017. [Online]. Available: https://catalogofbias.org/

[275] J. Kruger and D. Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of personality and social psychology*, vol. 77 6, pp. 1121–34, 1999.

[276] M. Delgado-Rodriguez, "Bias," *Journal of Epidemiology & Community Health*, vol. 58, no. 8, pp. 635–641, Aug. 2004. [Online]. Available: https://jech.bmj.com/lookup/doi/10.1136/jech.2003.008466

[277] P. M. Lukacs, K. P. Burnham, and D. R. Anderson, "Model selection bias and freedman's paradox," *Annals of the Institute of Statistical Mathematics*, vol. 62, pp. 117–125, 2009.

[278] K. Lerman and T. Hogg, "Leveraging Position Bias to Improve Peer Recommendation," *PLOS ONE*, vol. 9, no. 6, p. e98914, Jun. 2014, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098914

[279] A. Kaplan, *The conduct of inquiry*, A. Kaplan, Ed. Somerset, NJ: Transaction, Apr. 1998.

[280] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," *arXiv:1403.7400 [cs.SI]*, 2014.

[281] B. Goodman and S. Flaxman, "European Union regulations on algorithmic

decision-making and a ”right to explanation”,” *AIMag*, vol. 38, no. 3, pp. 50–57, Oct. 2017, arXiv: 1606.08813. [Online]. Available: http://arxiv.org/abs/1606.08813