Draft NIST Special Publication 1270

A Proposal for Identifying and Managing Bias in Artificial Intelligence

Reva Schwartz
Leann Down
Adam Jonas
Elham Tabassi

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

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

Draft NIST Special Publication 1270

A Proposal for Identifying and Managing Bias within Artificial Intelligence

Reva Schwartz
National Institute of Standards and Technology
Information Technology Laboratory

Leann Down
Adam Jonas
Parenthetic, LLC

Elham Tabassi
National Institute of Standards and Technology
Information Technology Laboratory

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

June 2021

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 (Draft)
Natl. Inst. Stand. Technol. Spec. Publ. 1270 (Draft),
30 pages (June 2021)

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

Organizations are encouraged to review this draft publication during the public comment period and provide feedback to NIST.

Public comment period: June 21 – September 10, 2021

National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, Maryland 20899-2000 70 Email: ai-bias@list.nist.gov

Abstract

NIST contributes to the research, standards, evaluation, and data required to advance the development and use of trustworthy artificial intelligence (AI) to address economic, social, and national security challenges and opportunities. Working with the AI community, NIST has identified the following 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. Mitigation of risk derived from bias in AI-based products and systems is a critical but still insufficiently defined building block of trustworthiness. This report proposes a strategy for managing AI bias, and describes types of bias that may be found in AI technologies and systems. The proposal is intended as a step towards consensus **standards** and a **risk-based framework** for trustworthy and responsible AI. The document, which also contains an alphabetical glossary that defines commonly occurring biases in AI, contributes to a fuller description and understanding of the challenge of harmful bias and ways to manage its presence in AI systems.

Key words

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

Table of Contents

Acknowledgments

The authors wish to thank the many people who assisted with the development of this document, including our NIST colleagues, and the many academic and technical reviewers who took the time to provide their valuable feedback.

Audience

The main audience for this document is researchers and practitioners in the field of trustworthy and responsible artificial intelligence. Researchers will find this document useful for understanding a view of the challenge of bias in AI, and as an initial step toward the development of standards and a risk framework for building and using trustworthy AI systems. Practitioners will benefit by gaining an understanding about bias in the use of AI systems.

Trademark Information

Note to Reviewers

As described throughout this report, one goal for NIST's work in trustworthy AI is the development of a **risk management framework** and accompanying **standards**. To make the necessary progress towards that goal, NIST intends to carry out a variety of activities in 2021 and 2022 in each area of the core building blocks of trustworthy AI (accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security (resilience), and mitigation of harmful bias). This will require a concerted effort, drawing upon experts from within NIST and external stakeholders. NIST seeks additional collaborative feedback from members of the research, industry, and practitioner community throughout this process. All interested parties are encouraged to please submit comments about this draft report, and the types of activities and events which would be helpful, via the public comment process described on page 3 of this document. There will also be opportunities for engaging in discussions about and contributing to development of key practices and tools to manage Bias in AI. Please look for announcements for webinars, call for position papers, and request for comment on NIST document(s).

1. Introduction

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).

In August 2019, fulfilling an assignment in an Executive Order[1] on AI, NIST released "A Plan for Federal Engagement in Developing Technical Standards and Related Tools." [100] 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 and enhance the trustworthiness of AI systems. Working with the AI community, NIST has identified the following 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.

This paper, *A Proposal for Identifying and Managing Bias in Artificial Intelligence*, has been developed to advance methods to understand and reduce harmful forms of AI bias. It is one of a series of documents and workshops in the pursuit of a **framework for trustworthy and responsible AI.**

While AI has significant potential as a transformative technology, it also poses inherent risks. One of those risks is bias. Specifically, how the presence of bias in automated systems can contribute to harmful outcomes and a public lack of trust. Managing bias is a critical but still insufficiently developed building block of trustworthiness.

The International Organization for Standardization (ISO) defines bias in statistical terms: "the degree to which a reference value deviates from the truth" [67]. This deviation from the truth can be either positive or negative, it can contribute to harmful or discriminatory outcomes or it can even be beneficial. From a societal perspective, bias is often connected to values and viewed through the dual lens of differential treatment or disparate impact, key legal terms related to direct and indirect discrimination, respectively.

Not all types of bias are negative, and there many ways to categorize or manage bias; this report focuses on biases present in AI systems that can lead to harmful societal outcomes. These harmful biases affect people's lives in a variety of settings by causing disparate impact, and discriminatory or unjust outcomes. The presumption is that bias is present throughout AI systems, the challenge is identifying, measuring, and managing it. Current approaches tend to classify bias by type (i.e.: statistical, cognitive), or use case and industrial sector (i.e.: hiring, health care, etc.), and may not be able to provide the broad perspective required for effectively managing bias as the context-specific phenomenon it is. This document attempts to bridge that

[1]https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence

gap and proposes an approach for managing and reducing the impacts of harmful biases[2] *across* contexts. The intention is to leverage key locations within stages of the AI lifecycle for optimally identifying and managing bias. As NIST develops a framework and standards in this area, the proposed approach is a starting point for community-based feedback and follow-on activities related to bias and its role in trustworthy AI.

2. The Challenge Posed by Bias in AI Systems

The proliferation of modeling and predictive approaches based on data-driven and machine learning techniques has helped to expose various social biases baked into real-world systems, and there is increasing evidence that the general public has concerns about the risks of AI to society. Distrust in AI can manifest itself through a belief that biases may be automated within these technologies, and can perpetuate harms more quickly, extensively, and systematically than human and societal biases on their own. Human decisions based on automated and predictive technology are often made in settings such as hiring or criminal justice, and can create harmful impacts and amplify and accelerate existing social inequities or, at minimum, perceptions of inequities. While it's unlikely that technology exhibiting "zero risk" can be developed, managing and reducing the impacts of harmful biases in AI is possible and necessary.

Public attitudes about AI technology suggest that, while often depending on the application, most Americans are unaware when they are interacting with AI enabled tech [53] but feel there needs to be a "higher ethical standard" than with other forms of technologies [76]. This mainly stems from the perceptions of fear of loss of control and privacy [47,125,133,137]. Certainly, there is no shortage of examples where bias in some aspect of AI technology and its use has caused harm and negatively impacted people's lives, such as in hiring [5,12,16,17,36,62,118], health care [46,52,55,59,83,88,103,122,123], and criminal justice [7,20,29,41,44,56,66,74,75,78,87, 140,142]. Indeed, there are many instances in which the deployment of AI technologies have been accompanied by concerns of whether and how societal biases are being perpetuated or amplified [3,10,14,15,22,24,34,42,45,61,102,105,108,116,126,139].

Since AI systems are deployed across various contexts, the associated biases that come with their use create harm in context-specific ways. This proliferation of AI bias into an ever-increasing list of settings makes it especially difficult to develop overarching guidance or mitigation techniques. A confounding factor is that it is especially difficult to predict where and how AI systems will be used. A current approach to the challenge of AI bias is to tackle a given use case where a particularly prevalent type of bias resides. This ad-hoc strategy is difficult to scale, and is unlikely to achieve what is required for building systems that the public can trust. Instead of viewing the challenge of AI bias within a given context or use case, a broader perspective can strike the problem of AI bias where it might be easiest to manage – within the design, development, and use of AI systems.

There are specific conditional traits associated with automation that exacerbate distrust in AI tools. One major purpose, and a significant benefit, of automated technology is that it can make sense of information more quickly and consistently than humans. There have long been two common assumptions about the rise and use of automation: it could make life easier [137] and

[2] For the purpose of this document the term "managing bias" will be used to refer to approaches for managing, reducing or mitigating bias.

also create conditions that reduce (or eliminate) biased *human* decision making and bring about a more equitable society [78]. These two tenets have led to the deployment of automated and predictive tools within trusted institutions and high-stake settings. While AI can help society achieve significant benefits, the convenience of automated classification and discovery within large datasets may come with a potentially significant downside. As these tools proliferate across our social systems, there has been increased interest in identifying and mitigating their harmful impacts.

The difficulty in characterizing and managing AI bias is exemplified by systems built to model concepts that are only partially observable or capturable by data. Without direct measures for these often highly complex considerations, AI development teams often use proxies. For example, for “criminality,” a measurable index, or construct, might be created from other information, such as past arrests, age, and region. For “employment suitability,” an AI algorithm might rely on time in prior employment, previous pay levels, education level, participation in certain sports [115], or distance from the employment site [51] (which might disadvantage candidates from certain neighborhoods).

There are many challenges that come with this common practice (see [89] for a thorough review). One challenge rests on the reality that decisions about which data to use for these indices are often made based on what is available or accessible, rather than what might be most suitable - but difficult or impossible to utilize [49]. Relatedly, instead of identifying specific questions of interest *first*, researchers, developers, and practitioners may “go where the data is” and adapt their questions accordingly [130]. Data can also differ significantly between what is collected and what occurs in the real world [71,72,109]. For example, responses to online questionnaires are from a specific sampling of the kinds of people who are online, and therefore leaves out many other groups. Data representing certain societal groups may be excluded in the training datasets used by machine learning applications [40]. And, datasets used in natural language processing often differ significantly from their real-world applications [113] which can lead to discrimination [128] and systematic gaps in performance.

Even if datasets are reflective of the real world, they may still exhibit entrenched historical and societal biases, or improperly utilize protected attributes. (Federal laws and regulations have been established to prohibit discrimination based on grounds such as gender, age, and religion.) Simply excluding these explicit types of attributes will not remedy the problem, however, since they can be inadvertently inferred in other ways (for example, browsing history), and still produce negative outcomes for individuals or classes of individuals [12]. So, 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.

Additionally, for much of the public, AI is not necessarily something with which they directly interact, and systems' algorithmic assumptions may not be transparent to them. Nevertheless, many people are affected or used as inputs by AI technologies and systems. This can happen when an individual applies for a loan [136], college [48], or a new apartment [77]. Historical, training data, and measurement biases are “baked-in” to the data used in the algorithmic models underlying those types of decisions. Such biases may produce unjust outcomes for racial and ethnic minorities in areas such as criminal justice [7,41,56,74,75,78,87,140,142], hiring [4,5,12,16,17,36,118,119], and financial decisions [13,65].

Another cause for distrust may be due to an entire class of untested and/or unreliable algorithms deployed in decision-based settings. Often a technology is not tested – or not tested extensively – before deployment, and instead deployment may be used *as* testing for the technology. An example is the rush to deploy systems during the COVID pandemic that have turned out to be methodologically flawed and biased [117,124,141]. There are also examples from the literature which describe technology that is based on questionable concepts, deceptive or unproven practices, or lacking theoretical underpinnings [2,9,13,30,33,62,129,141]. The broad consensus of the literature is that systems meant for decision making or predictive scenarios should demonstrate validity and reliability under the very specific setting in which it is intended to be deployed (hiring purposes, risk assessments in the criminal justice system, etc.). The decisions based on these algorithms affect people's lives in significant ways, and it is appropriate to expect protections in place to safeguard from certain systems and practices. The public's cautious opinions toward AI [138] might turn increasingly negative if *new* technologies appear which are based on the same approaches that have already contributed to systematic and well-documented societal harms.

To summarize the problem, there are many reasons for potential public distrust of AI related to bias in systems. These include:

- The use of datasets and/or practices that are inherently biased and historically contribute to negative impacts
- Automation based on these biases placed in settings that can affect people's lives, with little to no testing or gatekeeping
- Deployment of technology that is either not fully tested, potentially oversold, or based on questionable or non-existent science causing harmful and biased outcomes

Identifying and working to manage these kinds of bias can mitigate concerns about trustworthiness for in-place and in-development AI technologies and systems. An effective approach will likely need to be one that is not segmented by use case, but works across contexts.

Improving trust in AI systems can be advanced by putting mechanisms in place to reduce harmful bias in both deployed systems *and* in-production technology. Such mechanisms will require features such as a common vocabulary, clear and specific principles and governance approaches, and strategies for assurance. For the most part, the standards for these mechanisms and associated performance measurements still need to be created or adapted. The goal is not "zero risk," but to manage and reduce bias in a way that contributes to more equitable outcomes that engender public trust. These challenges are intertwined in complex ways and are unlikely to be addressed with a singular focus on one factor or within a specific use or industry.

3. Approach

In the lead-up to this report, the authors sought to capture common themes about the many ways bias is defined and categorized in AI technology. This was accomplished through a literature review, discussions with leaders in the field, a NIST-hosted workshop on bias in AI[3], and the evaluation of prominent topics across the broader AI research community. This work is not without precedent; there are previous attempts to define and classify AI bias [26,35,64,68,69,91,94,95,98,106,127].

The literature review consisted of a total of 313 articles, books, reports, and news publications about AI bias[4] from a variety of perspectives. In the survey of the literature, we identified a list of prominent biases present in AI that are contributors to societal harms. This list and accompanying definitions are presented in an alphabetical glossary in Appendix A.

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 considers AI within the social system it operates [3,18,19,31,34,40,41,43,71,97,118,120,134]. Taking social factors into consideration is necessary for achieving trustworthy AI, and 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 [11,49]. Such a change in perspective will require working with new stakeholders and developing guidance for effectively engaging social factors within a technical perspective. A key factor in this area is the many ways in which institutions indirectly drive the design and use of AI. Also, while AI practices may not intend to contribute to inequality or other negative forms of bias, there are always complex social factors that may be overlooked, especially since biases play out in context-specific ways and may not be captured or understood within one setting.

Whether statistical or societal, bias continues to be a challenge for researchers and technology developers seeking to develop and deploy trustworthy AI applications. How bias and trust interrelate is a key societal question, and understanding it will be paramount to improving acceptance of AI systems. A consistent finding in the literature is the notion that trust can improve if the public is able to interrogate systems and engage with them in a more transparent manner. Yet, 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" [80]. Creating such an authority requires standard practices, metrics, and norms. NIST has experience in creating standards and databases, and has been evaluating the algorithms used in biometric technologies since the 1960s. With the development of privacy and cybersecurity frameworks [99,101], NIST has helped organizations manage risks of the digital environment, and, through a series of reports and workshops, intends to contribute to a similar collaborative approach for managing AI trustworthiness as part of broader stakeholder efforts.

4. Identifying and Managing Bias in Artificial Intelligence

Improving trust in AI by mitigating and managing bias starts with identifying a structure for how it presents within AI systems and uses. We propose a three-stage approach derived from the AI

---

[3] For more information about this workshop see https://www.nist.gov/news-events/events/2020/08/bias-ai-workshop, and Appendix B of this document.

[4] The full bibliographic survey can be found at https://www.nist.gov/system/files/documents/2021/03/26/20210317_NIST%20AI_Bibliography.pdf

lifecycle, to enable AI designers and deployers to better relate specific lifecycle processes with the types of AI bias, and facilitate more effective management of it. Organizations that design and develop AI technology use the AI lifecycle to keep track of their processes and ensure delivery of high-performing functional tools - but not necessarily to identify harms or manage them. Currently, there is no single global or industrial AI lifecycle standard, but many versions used across multiple sectors and regions with a range of stages. The approach for identifying and managing AI bias proposed in this report is adapted from current versions of the AI lifecycle[5], and consists of three distinct stages, and presumed accompanying stakeholder groups. This approach is a starting point and NIST seeks feedback about its viability and implementation.

1. **PRE-DESIGN**: where the technology is devised, defined and elaborated
2. **DESIGN AND DEVELOPMENT**: where the technology is constructed
3. **DEPLOYMENT:** where technology is used by, or applied to, various individuals or groups.

Figure 1: A three-stage approach for managing AI bias

The following sub-sections provide key considerations and examples that highlight how statistical biases present across various stages of AI applications; and reflect and interact with the many human cognitive and societal biases that are inherent in the data, modeling, decision making, and practical processes associated with the use of AI systems across sectors and contexts.

[5] The following AI lifecycles were utilized as key guidance for this report: Centers of Excellence (CoE) at the US 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 https://www.iso.org/standard/81118.html)

PRE-DESIGN STAGE

Problem formulation and decision making

AI products start in the pre-design stage, where planning, problem specification, background research, and identification and quantification 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 benefitting from a technology solution. Since many of the downstream processes hinge on decisions from this stage, there is a lot of pressure here to "get things right." 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 individual and group heuristics and limited points of view, affect later stages and decisions in complex ways, and lead to biased outcomes [12,31,43,72,109,120]. This is a key juncture where well-developed guidance, assurance, and governance processes can assist business units and data scientists to collaboratively integrate processes that reduce bias without being cumbersome or blocking progress.

Operational settings and unknown impacts

Current assumptions in AI development often revolve around the idea of technological solutionism – the perception that technology will lead to only *positive* solutions. This perception, often combined with a singular focus on tool optimization, can be at odds with operational scenarios, increasing the difficulty for the practitioners who have to make sense of tool output – often in high stakes settings [96]. What seems like a good idea for how a given dataset can be utilized in a specific use case might be perceived differently by the systems' end users or those affected by the systems' decisions. It is an obvious risk to build algorithmic-based decision tools for settings already known to be discriminatory. Yet, awareness of which conditions will lead to disparate impact or other negative outcomes is not always apparent in pre-design, and can be easily overlooked once in production.

Overselling tool capabilities and performance

Whether unconscious or unintentional, pre-design is often where decisions are made that can inadvertently lead to harmful impact, or be employed to extremely negative societal ends. By not addressing the possibility of optimistic and potentially inflated expectations related to AI systems, risk management processes could fail to communicate and set reasonable limits related to mitigating such potential harms. In extreme cases, with tools or apps that are fraudulent, pseudoscientific, prey on the user, or generally exaggerate claims, the goal should not be to ensure tools 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.

Other problems that can occur in pre-design include poor problem framing, basing technology on spurious correlations from data-driven approaches, failing to establish appropriate underlying causal mechanisms, or generally technically flawed [22,34,40,52,54,89,102,110]. In such cases (often termed "fire, ready, aim"), the solution may not be mitigation, but rather, rejection of the system or the way in which the perceived underlying problem is framed. These types of scenarios may reinforce public distrust of AI technology as systems that are untested or technically flawed can also contribute to bias. Technology designed for use in high-stakes settings requires extensive testing to demonstrate valid and reliable performance [58,112].

Practices

There is currently momentum for AI researchers to include statements about the potential societal impacts [114] when submitting their work to journals or conferences. Identifying and addressing potential biases early in the problem formulation process is an important step in this process. It is also complicated by the role of power and decision making [96]. 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) [32]. These kinds of practices can lead to a more thorough evaluation of the broad societal impacts of technology-based tools across the three stages. 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. 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 or tools 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 can significantly contribute to positive outcomes.

Real-world example

There are many examples of bias from the real world where practices in the problem formulation stage may have combined with lack of understanding of downstream impacts. For example, the Gender Shades facial recognition evaluation project [24] describes the poor performance of facial recognition systems when trying to detect face types (by gender and skin type) that are not present in the training data. This is an example of representation bias – a type of sampling bias that pre-dates AI - where trends estimated for one population are inappropriately generalized to data collected from another population. This biased performance was not identified by the teams that designed and built the facial recognition systems, but instead by researchers evaluating the systems' performance in different conditions. It is during the pre-design stage where these kinds of implicit decisions are made about what constitutes a "valid face," and non-representative datasets are selected. Additionally, representation bias can lead to bigger problems and other biases in later stages of the AI lifecycle, an issue referred to as "error propagation," that can eventually lead to biased outcomes [90]. Improving pre-design practices to ensure more inclusive representation can help to broaden the larger teams' perspectives about what is considered relevant or valid.

DESIGN AND DEVELOPMENT STAGE

This stage of the AI lifecycle is where modeling, engineering and validation take place. The stakeholders in this stage tend to include software designers, engineers, and data scientists who carry out risk management techniques in the form of algorithmic auditing and enhanced metrics for validation and evaluation.

Optimization over context

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 [50], selecting models based solely on accuracy is not necessarily the best approach for bias reduction. Not taking context into consideration during model selection can lead to biased results for sub-populations (for example, disparities in health care delivery). Relatedly, tools 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 [7]- can lead to biased outcomes. This type of bias, known as ecological fallacy, occurs when an inference is made about an individual based on their membership within a group (for example, basing college admissions decisions on an individual's race) [48]. These unintentional weightings of certain factors can cause algorithmic results that exacerbate and reinforce societal inequities. The surfacing of these inequities is a kind of positive "side effect" of algorithmic modeling, enabling the research community to discover them and develop methods for managing them.

Practices

During modeling tasks in this stage, it may become apparent that algorithms are biased or will contribute to disparate impacts if deployed. In such cases the technology can be taken out of production. But this kind of awareness and remedy is likely to take place only in certain settings or industries, with well-defined procedures and clear lines of accountability. Unfortunately, not all tools are deployed in such settings – and capturing the wide array of use cases and scenarios is particularly difficult. It is also notable that, depending on the industry or use case, AI is typically marketed as an easy solution that does not necessarily require extensive support. But the notion that AI requires extensive monitoring belies the reality that AI can be both easy to use *and* should be used with extreme caution [96].

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. For example, "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 [60]. Requiring AI practitioners to defend their techniques can incentivize new ways of thinking, stimulate improved practices, and help create change in approaches by individuals and organizations [96]. To better identify and mitigate organizational factors which can contribute to bias, experts also suggest the use of algorithmic decision-making tools for specific, well-defined use cases, and *not beyond* those use cases (a factor that will be discussed more in-depth in the section about deployment). Additionally, researchers also 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 tool output [111].

Real-world example

One real-world case of a biased outcome that may have been manageable at the design and development stage is the university admissions algorithm GRADE [135], which was shown to produce biased enrollment decisions for incoming PhD students [25]. Without ground truth for what constitutes a "good fit," a construct was developed using prior admission data. Once put into production, the model ended up being trained to do a different job than intended (also known as "target leakage"). Instead of assessing student quality, the model learned previous admissions officer decisions. Another issue is that candidate quality cannot be truly known until after the student matriculates. This case is a good example of data hubris, or "overstated claims that arise from big data analysis" [84]. This is particularly problematic when using data to "make causal claims from an inherently inductive method of pattern recognition" [19,84,89].

DEPLOYMENT STAGE

This stage is where users start to interact with the developed technology, and sometimes create unintended uses for it. The stakeholders in deployment are often the different types of end users who directly interact with technology tools for their profession. This includes operators, subject matter experts, humans-in-the-loop, and decision-makers who interpret output to make or support decisions.

Discriminatory impact

Since many AI-based tools can skip deployment to a specified expert end user, and are marketed to, and directly used by, the general public, the intended uses for a given tool are often quickly overcome by reality. Additionally, members of the public do not necessarily have to directly interact with technology to be affected by tool deployment. Individuals' data can be used for modeling (sometimes without their knowledge), and in decisions that can affect their lives based on factors such as where they live and work. For example, the algorithms used in ride hailing apps learned the landscape of low-income non-white neighborhoods and charged citizens who live there more for pick-up and drop-off, causing disparate impact [108]. This kind of systemic discriminatory pricing is perpetuated on the citizens of the neighborhood without their knowledge, whether they have and use the app or not, and due only to the fact that they live there.

Intended context vs. actual context

Once people start to interact with an AI system, early design and development decisions that were poorly or incompletely specified or based on narrow perspectives can be exposed. This leaves the process vulnerable to additive biases that are either statistical in nature or related to human decision making and behavior [109]. For example, by not designing to compensate for activity biases, algorithmic models may be built on data only from the most active users, likely creating downstream system activity that does not reflect the intended or real user population [1,8]. Basing system actions on an unrepresentative sample can have significant impact. For example, by not considering that STEM ads 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 [82].

The deployment stage also offers an interesting window into how perceptions and uses can differ based on the distance from the technology itself. In pre-design the focus and perceptions are about how technology can be designed to solve a question, market a product, or innovate in a new area. In design and development, the focus is on building, testing, and operationalizing the technology, typically with time to market and accuracy as the key criteria. And once the technology is deployed and used in different settings and for different purposes, we see perceptions turn to unintended use cases and even distrust. In one case of predictive analytics in university admissions, the operators of the receiving end of the tool output were the ones to sound the warning about race-based biases [79]. Although the study was based on a small number of participants, interviews with admissions officials suggest that "they didn't believe in the validity of the risk scores, they thought the scores depersonalized their interactions with students, and they didn't understand how the scores were calculated" [48].

The kinds of scenarios where experts utilize and rely upon automated results (like in the college admissions example), are highly complex and relatively understudied. One key issue is finding a configuration that enables a system to be used in a way that optimally leverages, instead of replaces, user expertise. This is often a significant challenge since domain experts and AI developers often lack a common vernacular, which can contribute to miscommunication and misunderstood capabilities. With the promise of more quantitative approaches, domain experts may tend to offload method validation to the AI system itself. End users may also subconsciously find ways to leverage those perceived "objective" results as cover for their biases [6,38,39]. On the system side, developer communities may presume method validation at a level that is not actually present. These kinds of loopholes can create conditions that operationalize technology that is not quite ready for use, especially in high-stakes settings [11,120].

<u>Contextual gaps lead to performance gaps</u>

The "distance from technology" can also contribute to different types of performance gaps. There are gaps in intention; these are gaps between what was originally intended in pre-design versus what is developed and between the AI product and how it is deployed. There also are gaps in performance based on those intention gaps. When an AI tool is designed and developed to be used in a specific setting and tested for use in near-laboratory conditions, there are clearer expectations about intended performance. Once the AI tool is deployed and goes "off-road," the original intent, idea, or impact assessment that was identified in pre-design can drift as the tool is repurposed and/or used in unforeseen ways.

Another important gap that contributes to bias relates to differences in interpretability requirements between users and developers. As previously discussed, the groups who invent and produce technology have specific intentions for its use and are unlikely to be aware of all the ways a given tool will be repurposed. There are individual differences in how humans interpret AI model output. When system designers do not take these differences into consideration it can contribute to misinterpretation of that output [21]. When these differences are combined with the societal biases found in datasets and human cognitive biases such as automation complacency (which is particularly relevant in the deployment stage), where end users may unintentionally "offload" their decisions to the automated tool - this can cause significant negative impacts.

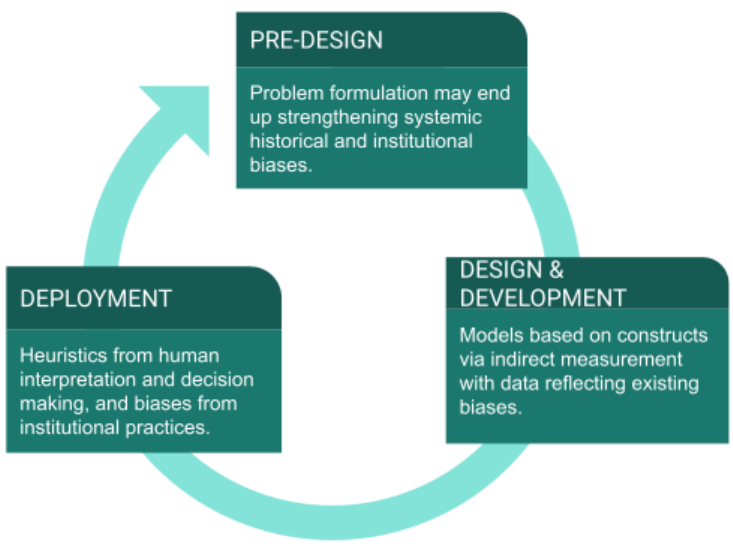
Practical improvements

One approach for managing bias risks associated with the gaps described above is deployment monitoring and auditing. Counterfactual fairness is a technique used by researchers to bridge the gaps between the laboratory and the post-deployment real world. The issue, as described in [81] is that "*If individuals in the training data have not already had equal opportunity, algorithms enforcing EO*[6] *will not remedy such unfairness*." Using the GRADE algorithm as an example, instead of using previous admission decisions as the predictor, the model would consider and seek to compensate for the various social biases that could impact a student's application. This happens by capturing "these social biases and make clear the implicit trade-off between prediction accuracy and fairness in an unfair world." Identifying standards of practice for implementing these types of risk management tools and techniques will be a focus of future activities.

Summary

In this section we have described the challenge of AI bias and proposed an approach for considering how to manage it through three stages modeled on AI development lifecycle. The section also shows that, while the type of bias and manner of presentation may differ, bias can occur across all of these stages. To summarize and help illustrate this point, the below figure shows an exemplar of how bias could present within each of the three stages.

Figure 2: Example of bias presentation in three stages modeled on the AI lifecycle.



[6] EO = equal opportunity

5. Conclusion and Next Steps

We have identified a few of the many ways that algorithms can create conditions for discriminatory decision making. In an effort to identify the technical requirements for cultivating trustworthy and responsible AI, this report suggests a three-stage approach for managing AI bias. This approach is intended to foster discussion about the path forward and collaborative development of **standards** and a risk-based **framework**. Rather than identifying and tackling specific biases within cases, this report suggests a need to address the context-specific nature of AI bias by associating applicable biases within specific stages modeled on the AI lifecycle for more effective management and mitigation. NIST is interested in obtaining feedback from the broader community about this proposed approach via public comment and a series of public events.

The broader AI research community, practitioners, and users all have many valuable insights and recommendations to offer in managing and mitigating bias. Identifying which techniques to include in a framework that seeks to promote trustworthiness and responsibility in AI requires an approach that is actively representative and includes a broad set of disciplines and stakeholders. This will allow interested parties to move forward with guidance that is effective *and* implementable, accurate, realistic, and fit for purpose. It has the potential to increase public trust and advance the development and use of beneficial AI technologies and systems. To that end, this report concludes:

- Bias is neither new nor unique to AI.
- The goal is not zero risk but rather, identifying, understanding, measuring, managing and reducing bias.
- Standards and guides are needed for terminology, measurement, and evaluation of bias.
- Bias reduction techniques are needed that are flexible and can be applied across contexts, regardless of industry.
- NIST plans to develop a framework for trustworthy and responsible AI with the participation of a broad set of stakeholders to ensure that standards and practices reflect viewpoints not traditionally included in AI development.
- NIST will collaboratively develop additional guidance for assurance, governance, and practice improvements as well as techniques for enhancing communication among different stakeholder groups.

To make the necessary progress towards the goal of trustworthy and responsible AI, NIST intends to act as a hub for the broader community of interest and to collaboratively engage with experts and other stakeholders as they address the challenges of AI. To that end, NIST will host a variety of activities in 2021 and 2022 in each area of the core building blocks of trustworthy AI (accuracy, explainability and interpretability, privacy, reliability, robustness, safety, and security (resilience), and bias).

6. Appendices

Appendix A: Glossary

The table below presents a glossary with a stand-alone definition for each term and accompanying reference(s). The goal and contribution of this glossary is to aggregate terms that are in common usage or relevance to AI bias. Definitions were selected based on either recently published papers from the AI bias community or seminal work in the area the term is most associated with. When multiple definitions of a bias 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.

Table 1: Bias Terminology. This table lists definitions with accompanying references for select biases in AI.

Bias type	Definition
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) [8].
Amplification bias	Arises when the distribution over prediction outputs is skewed in comparison to the prior distribution of the prediction target [85].
Annotator bias, Human reporting bias	When users rely on automation as a heuristic replacement for their own information seeking and processing [93].
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).
Behavioral bias	Systematic distortions in user behavior across platforms or contexts, or across users represented in different datasets [92,104].
Cognitive bias	Systematic errors in human thought based on a limited number of heuristic principles and predicting values to simpler judgmental operations [132].
Concept drift, 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.
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 [121].
Content production bias	Arises from structural, lexical, semantic, and syntactic differences in the contents generated by users [104].
Data generation bias	Arises from the addition of synthetic or redundant data samples to a dataset [73].

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 [127].
Detection bias	Systematic differences between groups in how outcomes are determined and may cause an over- or underestimation of the size of the effect [27].
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 [127].
Exclusion bias	When specific groups of user populations are excluded from testing and subsequent analyses [37].
Feedback loop bias	Effects that may occur when an algorithm learns from user behavior and feeds that behavior back into the model [121].
Funding bias	Arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study [91].
Historical bias	Arises when models are trained on past (potentially biased) decisions [72].
Inherited bias, Error propagation	Arises when tools that are built with machine learning are used to generate inputs for other machine learning algorithms. If the output of the tool is biased in any way, this bias may be inherited by systems using the output as input to learn other models [64].
Institutional bias, Systemic bias	A tendency for the procedures and practices of particular institutions to 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 simply following existing rules or norms. Institutional racism and institutional sexism are the most common examples [28].
Interpretation bias	A form of information processing bias that can occur when users interpret algorithmic outputs according to their internalized biases and views [121].
Linking bias	Arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users [104].
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.
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 [127].

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.
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 [1].
Population bias	Arises when statistics, demographics, and user characteristics differ between the original target population and the user population represented in the actual dataset or platform [91].
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 [8].
Ranking bias	The idea that top-ranked results are the most relevant and important and will result in more clicks than other results [8,86].
Sampling bias, 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 [91].
Selection bias	Bias that results from using nonrandomly selected samples to estimate behavioral relationships as an ordinary specification bias that arises because of a missing data problem [63].
Selective adherence	Decision-makers' inclination to selectively adopt algorithmic advice when it matches their pre-existing beliefs and stereotypes [6].
Societal bias	Ascribed attributes about social groups that are largely determined by the social context in which they arise and are an adaptable byproduct of human cognition [23].
Statistical bias	A systematic tendency for estimates or measurements to be above or below their true values. Note 1: Statistical biases arise from systematic as opposed to random error. Note 2: Statistical bias can occur in the absence of prejudice, partiality, or discriminatory intent [107].
Temporal bias	Bias that arises from differences in populations and behaviors over time [104,131].
Training data bias	Biases that arise from algorithms that are trained on one type of data and do not extrapolate beyond those data.
Uncertainty bias, Epistemic uncertainty	Arises when predictive algorithms favor groups that are better represented in the training data, since there will be less uncertainty associated with those predictions [57].
User interaction bias	Arises when a user imposes their own self-selected biases and behavior during interaction with data, output, results, etc. [8].

Appendix B: Collaborative Work

This report is based on a series of collaborative events, including a literature review, input from leaders in the field through ongoing discussions and a workshop, and a broad evaluation of the significant themes across the community of interest. Detailed information of these events is described below.

Literature review

During 2020, NIST implemented a broad review of materials from frequently-cited, shared, and cross-referenced pieces focused on bias within technologies that use artificial intelligence. This review incorporated content that described AI bias from a societal perspective, in existing technologies and development processes, and other factors that influence AI development, implementation, and/or adaptation. To ensure a cross-section of perspectives, literature was identified across a variety of publication types, including peer-reviewed journals, popular news media, books, organizational reports, conference proceedings, and presentations. Across publications, the literature review topics represent a wide range of stakeholder perspectives and challenges and current and future AI implementations.

Workshop on Bias in AI

Recognizing a lack of consensus regarding several fundamental concepts in identifying and understanding bias in AI, NIST convened a virtual workshop August 18, 2020 with experts, researchers, and stakeholders from a variety of organizations and sectors whose work focuses on the topic. The workshop consisted of panel discussions on data and algorithmic bias, followed by five contemporaneous breakout sessions. Notes from workshop organizers, facilitators, and scribes were reviewed for key takeaways and themes. Workshop participants suggested that forums and workshops like the one held on August 18 were important to maintaining awareness and alignment of current challenges and future solutions. Participants also referred to the long-term nature of this challenge. These key takeaways have been included and described throughout this report.

7. References

[1] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The Unfairness of Popularity Bias in Recommendation, ArXiv:1907.13286 [Cs]. (2019).

[2] B. Aguera y Arcas, M. Mitchell, A. Todorov, Physiognomy's new clothes, Medium. (2017).

[3] M. Aitken, E. Toreini, P. Carmichael, K. Coopamootoo, K. Elliott, 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. 7 (2020) 205395172090889.

[4] I. Ajunwa, The Paradox of Automation as Anti-Bias Intervention, Cardozo L. Rev. 41 (2020) 1671.

[5] I. Ajunwa, S. Friedler, C. Scheidegger, S. Venkatasubramanian, Hiring by Algorithm: Predicting and Preventing Disparate Impact, SSRN Electric Journal. (2016).

[6] S. Alon-Barkat, M. Busuioc, Decision-makers Processing of AI Algorithmic Advice: Automation Bias versus Selective Adherence, ArXiv:2103.02381 [Cs]. (2021).

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

[8] R. Baeza-Yates, Bias on the web, Commun. ACM. 61 (2018) 54–61.

[9] K. Bailey, Put Away Your Machine Learning Hammer, Criminality Is Not A Nail, Wired. (2016).

[10] J. Bajorek, Voice Recognition Still Has Significant Race and Gender Biases, Harvard Business Review. (2019).

[11] S. Barocas, A. Biega, B. Fish, J. Niklas, L. Stark, When not to design, build, or deploy, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2020: p. 695.

[12] S. Barocas, A. Selbst, Big Data's Disparate Impact, California Law Review. 104 (2016) 671–732.

[13] R. Bartlett, A. Morse, R. Stanton, N. Wallace, Consumer-Lending Discrimination in the FinTech Era, National Bureau of Economic Research, 2019.

[14] E. Bary, How artificial intelligence could replace credit scores and reshape how we get loans, Market Watch. (2018).

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

[16] M. Bogen, All the Ways Hiring Algorithms Can Introduce Bias, Harvard Business Review. (2019).

[17] M. Bogen, A. Rieke, Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias, Upturn. (2019).

[18] M. Boyarskaya, A. Olteanu, K. Crawford, Overcoming Failures of Imagination in AI Infused System Development and Deployment, ArXiv:2011.13416 [Cs]. (2020).

[19] D. Boyd, K. Crawford, CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon, Information, Communication & Society. 15 (2012) 662–679.

[20] S. Brayne, Enter the Dragnet, Logic Magazine. (2020).

[21] D. Broniatowski, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence, NIST, 2021.
[22] M. Broussard, Artificial Unintelligence: How Computers Misunderstand the World, MIT Press, 2018.
[23] R. Brown, S. Gaertner, Blackwell Handbook of Social Psychology: Intergroup Processes, John Wiley & Sons, 2008.
[24] J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Proceedings of Machine Learning Research, 2018: pp. 77–91.
[25] L. Burke, U of Texas will stop using controversial algorithm to evaluate Ph.D. applicants | Inside Higher Ed, Inside Higher Ed. (2020).
[26] A. Caliskan, J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science. 356 (2017) 183–186.
[27] Centre for Evidence-Based Medicine, Catalogue of Bias, Catalog of Bias. (2017).
[28] D. Chandler, R. Munday, A Dictionary of Media and Communication, Oxford University Press, 2011.
[29] A. Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, Big Data. 5 (2017) 153–163.
[30] Coalition for Critical Technology, Abolish the #TechToPrisonPipeline, (2020).
[31] S. Costanza-Chock, Design Justice, A.I., and Escape from the Matrix of Domination, Journal of Design and Science. (2018) 96c8d426.
[32] K. Crawford, Artificial Intelligence's White Guy Problem, The New York Times. (2016).
[33] K. Crawford, Time to regulate AI that interprets human emotions, Nature. 592 (2021) 167–167.
[34] C. Criado Perez, Invisible Women: Data Bias in a World Designed for Men, Abrams Press, 2019.
[35] D. Danks, A. London, Algorithmic Bias in Autonomous Systems, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 2017: pp. 4691–4697.
[36] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, Reuters. (2018).
[37] M. Delgado-Rodriguez, Bias, Journal of Epidemiology & Community Health. 58 (2004) 635–641.
[38] B.J. Dietvorst, J.P. Simmons, C. Massey, Algorithm aversion: people erroneously avoid algorithms after seeing them err, J Exp Psychol Gen. 144 (2015) 114–126.
[39] B.J. Dietvorst, J.P. Simmons, C. Massey, Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them, Management Science. 64 (2018) 1155–1170.
[40] C. D'Ignazio, L. Klein, Data Feminism, MIT Press, 2020.
[41] L. Dormehl, Algorithms Are Great and All, But They Can Also Ruin Lives, Wired. (2014).
[42] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness Through Awareness, ArXiv:1104.3913 [Cs]. (2011).
[43] M.C. Elish, S. Barocas, A. Plasek, K. Ferryman, The social & economic implications of artificial intelligence technologies in the near-term, AI Now, New York, 2016.

[44] EPIC, Algorithms in the Criminal Justice System: Risk Assessment Tools, Electronic Privacy Information Center, 2020.
[45] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.
[46] M. Evans, A.W. Mathews, New York Regulator Probes UnitedHealth Algorithm for Racial Bias, Wall Street Journal. (2019).
[47] E. Fast, E. Horvitz, Long-Term Trends in the Public Perception of Artificial Intelligence, in: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2017: p. 7.
[48] T. Feathers, Major Universities Are Using Race as a "High Impact Predictor" of Student Success – The Markup, The Markup. (2021).
[49] B. Fish, L. Stark, Reflexive Design for Fairness and Other Human Values in Formal Models, ArXiv:2010.05084 [Cs]. (2020).
[50] J.Z. Forde, A.F. Cooper, K. Kwegyir-Aggrey, C. De Sa, M. Littman, Model Selection's Disparate Impact in Real-World Deep Learning Applications, ArXiv:2104.00606 [Cs]. (2021).
[51] G. Friedman, T. McCarthy, Employment Law Red Flags in the Use of Artificial Intelligence in Hiring, American Bar Association. (2019).
[52] H. Fry, Hello world: being human in the age of algorithms, WW Norton & Company, 2018.
[53] S. Furman, J. Haney, Is My Home Smart or Just Connected?, in: International Conference on Human-Computer Interaction, Springer, Cham, 2020: pp. 273–287.
[54] D. Gaffney, N. Matias, Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus, PLOS ONE. 13 (2018) e0200162.
[55] M. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data, JAMA Intern Med. 178 (2018) 1544–1547.
[56] S. Goel, R. Shroff, J.L. Skeem, C. Slobogin, The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment, SSRN Journal. (2018).
[57] B. Goodman, S. Flaxman, European Union regulations on algorithmic decision-making and a "right to explanation," AIMag. 38 (2017) 50–57.
[58] P. Grother, M. Ngan, K. Hanaoka, Face recognition vendor test part 3: demographic effects, National Institute of Standards and Technology, Gaithersburg, MD, 2019.
[59] E. Guo, K. Hao, This is the Stanford vaccine algorithm that left out frontline doctors, MIT Technology Review. (2020).
[60] P. Hall, N. Gill, B. Cox, Responsible machine learning: Actionable strategies for mitigating risks and driving adoption, O'Reilly Media Inc., Sebastopol, CA, 2020.
[61] M. Hardt, E. Price, N. Srebro, Equality of Opportunity in Supervised Learning, ArXiv:1610.02413 [Cs]. (2016).
[62] E. Harlen, O. Schnuck, Objective or Biased, Bayerischer Randfunk. (2021).
[63] J. Heckman, Sample Selection Bias as a Specification Error, Econometrica. 47 (1979) 153–161.
[64] T. Hellström, V. Dignum, S. Bensch, Bias in Machine Learning -- What is it Good for?, ArXiv:2004.00686 [Cs]. (2020).
[65] M. Henry-Nickie, How artificial intelligence affects financial consumers, Brookings. (2019).

[66] K. Hill, Flawed Facial Recognition Leads To Arrest and Jail for New Jersey Man, The New York Times. (2020).
[67] ISO, Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability, 2006.
[68] ISO/IEC, Information technology — Vocabulary, International Organization for Standardization, Geneva, Switzerland, 2015.
[69] ISO/IEC, Information technology — Big data — Overview and vocabulary, International Organization for Standardization, Geneva, Switzerland, 2019.
[70] IT Modernization CoE, CoE Guide to AI Ethics, General Services Administration, n.d.
[71] A. Jacobs, S.L. Blodgett, S. Barocas, H. Daumé, H. Wallach, The meaning and measurement of bias: lessons from natural language processing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2020: p. 706.
[72] A. Jacobs, H. Wallach, Measurement and Fairness, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. (2021) 375–385.
[73] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, S.-L. Kim, Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data, ArXiv:1811.11479 [Cs, Stat]. (2018).
[74] J.E. Johndrow, K. Lum, An algorithm for removing sensitive information: Application to race-independent recidivism prediction, Ann. Appl. Stat. 13 (2019) 189–220.
[75] F. Kamiran, A. Karim, S. Verwer, 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, IEEE, Brussels, Belgium, 2012: pp. 370–377.
[76] A. Kerr, M. Barry, J. Kelleher, Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance, Big Data & Society. 7 (2020) 2053951720915939.
[77] L. Kirchner, M. Goldstein, Access Denied: Faulty Automated Background Checks Freeze Out Renters – The Markup, The Markup. (2020).
[78] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions, The Quarterly Journal of Economics. 133 (2018) 237–293.
[79] S. Klempin, M. Grant, M. Ramos, Practitioner Perspectives on the Use of Predictive Analytics in Targeted Advising for College Students, Community College Research Center, 2018.
[80] B. Knowles, J. Richards, The Sanction of Authority: Promoting Public Trust in AI, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2021: pp. 262–271.
[81] M.J. Kusner, J.R. Loftus, C. Russell, R. Silva, Counterfactual Fairness, ArXiv:1703.06856 [Cs, Stat]. (2018).
[82] A. Lambrecht, C. Tucker, Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads, SSRN Journal. 65 (2019) 2966–2981.
[83] H. Ledford, Millions of black people affected by racial bias in health-care algorithms, Nature. 574 (2019) 608–609.
[84] E. Lee, A. Yee, Toward Data Sense-Making in Digital Health Communication Research: Why Theory Matters in the Age of Big Data, Front. Commun. 5 (2020) 1–11.

[85] K. Leino, E. Black, M. Fredrikson, S. Sen, A. Datta, Feature-Wise Bias Amplification, ArXiv:1812.08999 [Cs, Stat]. (2019).
[86] K. Lerman, T. Hogg, Leveraging Position Bias to Improve Peer Recommendation, PLOS ONE. 9 (2014) e98914.
[87] A. Liptak, Sent to Prison by a Software Program's Secret Algorithms, The New York Times. (2017).
[88] T. Maddox, J. Rumsfeld, P. Payne, Questions for Artificial Intelligence in Health Care, JAMA. 321 (2019) 31–32.
[89] M. Malik, A Hierarchy of Limitations in Machine Learning, ArXiv:2002.05193 [Cs, Econ, Math, Stat]. (2020).
[90] G. McGraw, H. Figueroa, V. Shepardson, R. Bonett, An architectural risk analysis of machine learning systems: Toward more secure machine learning, Berryville Institute of Machine Learning, Clarke County, VA, 2020.
[91] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, ArXiv:1908.09635 [Cs]. (2019).
[92] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, 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, 2016: pp. 259–268.
[93] I. Misra, C.L. Zitnick, M. Mitchell, 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), IEEE, Las Vegas, NV, USA, 2016: pp. 2930–2939.
[94] M. Mitchell, Artificial Intelligence: A Guide for Thinking Human, Farrar, Straus, and Giroux, 2019.
[95] S. Mitchell, J. Shadlen, Mirror mirror: Reflections on quantitative fairness. Shira Mitchell: Statistician, Shira Mitchell: Statistician. (2020).
[96] E. Moss, J. Metcalf, Ethics Owners, Data & Society. (2020).
[97] E. Moss, J. Metcalf, High Tech, High Risk: Tech Ethics Lessons for the COVID-19 Pandemic Response, Patterns. 1 (2020) 100102.
[98] D.K. Mulligan, J.A. Kroll, N. Kohli, R.Y. Wong, This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology, Proc. ACM Hum.-Comput. Interact. 3 (2019) 1–36.
[99] NIST, Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1, National Institute of Standards and Technology, Gaithersburg, MD, 2018.
[100] NIST, U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools, National Institute of Standards and Technology, 2019.
[101] NIST, NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0, National Institute of Standards and Technology, Gaithersburg, MD, 2020.
[102] S.U. Noble, Algorithms of oppression: How search engines reinforce racism, NYU Press, 2018.
[103] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science. 366 (2019) 447–453.
[104] A. Olteanu, C. Castillo, F. Diaz, E. Kıcıman, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Front. Big Data. 2 (2019) 13.

[105] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Broadway Books, 2017.
[106] Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence, OECD Legal Instruments. (2019).
[107] Organization of Scientific Area Committees for Forensic Science, OSAC Preferred Terms, OSAC Preferred Terms. (2021).
[108] A. Pandey, A. Caliskan, Iterative Effect-Size Bias in Ridehailing: Measuring Social Bias in Dynamic Pricing of 100 Million Rides, ArXiv:2006.04599 [Cs]. (2020).
[109] S. Passi, S. Barocas, Problem Formulation and Fairness, Proceedings of the Conference on Fairness, Accountability, and Transparency. (2019) 39–48.
[110] J. Pfeffer, K. Mayer, F. Morstatter, Tampering with Twitter's Sample API, EPJ Data Sci. 7 (2018) 50.
[111] S. Picard, M. Watkins, M. Rempal, A. Kerodal, Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness, Center for Court Innovation, 2020.
[112] A. Picchi, Job hunters face a new hurdle: Impressing AI, CBS News. (2020).
[113] B. Plank, What to do about non-standard (or non-canonical) language in NLP, ArXiv:1608.07836 [Cs]. (2016).
[114] C. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, A. Dafoe, Institutionalizing ethics in AI through broader impact requirements, Nature Machine Intelligence. 3 (2021) 104–110.
[115] M. Raghavan, S. Barocas, Challenges for mitigating bias in algorithmic hiring, Brookings. (2019).
[116] J. Redden, The Harm That Data Do, Scientific American. (2018).
[117] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. Rudd, E. Sala, C.-B. Schönlieb, Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nature Machine Intelligence. 3 (2021) 199–217.
[118] J. Sanchez-Monedero, L. Dencik, 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]. (2020).
[119] H. Schellmann, Auditors are testing hiring algorithms for bias, but big questions remain, MIT Technology Review. (2021).
[120] A.D. Selbst, D. Boyd, S.A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and Abstraction in Sociotechnical Systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, ACM Press, Atlanta, GA, USA, 2019: pp. 59–68.
[121] S. Silva, M. Kenney, Algorithms, platforms, and ethnic bias, Commun. ACM. 62 (2019) 37–39.
[122] T. Simonite, How an Algorithm Blocked Kidney Transplants to Black Patients, Wired. (2020).
[123] M. Singh, K. Ramamurthy, Understanding racial bias in health using the Medical Expenditure Panel Survey data, ArXiv:1911.01509 [Cs, Stat]. (2019).
[124] J. Sipior, Considerations for development and use of AI in response to COVID-19, International Journal of Information Management. 55 (2020) 102170.
[125] A. Smith, M. Anderson, Automation in Everyday Life, Pew Research Center, 2017.

[126] M. Specia, Siri and Alexa Reinforce Gender Bias, U.N. Finds, The New York Times. (2019).
[127] H. Suresh, J. Guttag, A Framework for Understanding Unintended Consequences of Machine Learning, ArXiv:1901.10002 [Cs, Stat]. (2020).
[128] Y.C. Tan, L.E. Celis, Assessing Social and Intersectional Biases in Contextualized Word Representations, ArXiv:1911.01485 [Cs, Stat]. (2019).
[129] M. Tobin, L. Matsakis, China is home to a growing market for dubious "emotion recognition," Rest of World. (2021).
[130] R. Tromble, Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age, Social Media + Society. 7 (2021) 2056305121988929.
[131] Z. Tufekci, Big questions for social media big data: Representativeness, validity and other methodological pitfalls, ArXiv:1403.7400 [Cs.SI]. (2014).
[132] A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases, Science. 185 (1974) 1124–1131.
[133] W. Ware, Records, Computers and the Rights of Citizens, RAND Corporation, Santa Monica, CA, 1973.
[134] A.L. Washington, 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, ACM, Barcelona Spain, 2020: pp. 230–240.
[135] A. Waters, R. Miikkulainen, GRADE: Machine Learning Support for Graduate Admissions, AIMag. 35 (2014) 64–64.
[136] M. Weber, M. Yurochkin, S. Botros, V. Markov, Black Loans Matter: Distributionally Robust Fairness for Fighting Subgroup Discrimination, ArXiv:2012.01193 [Cs]. (2020).
[137] D. West, Brookings survey finds worries over AI impact on jobs and personal privacy, concern U.S. will fall behind China, Brookings. (2018).
[138] D. West, Brookings survey finds divided views on artificial intelligence for warfare, but support rises if adversaries are developing it, Brookings. (2018).
[139] D. West, J. Allen, Turning Point, Brookings. (2020).
[140] R. Wexler, When a Computer Program Keeps You in Jail, The New York Times. (2017).
[141] L. Wynants, B.V. 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, M. Smeden, Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ. 369 (2020) m1328.
[142] State v. Loomis, 2016.