Pre-workshop brief

NIST Workshop on Bias in AI
August 18, 2020
9:00am - 5:00pm EDT

The NIST Workshop on Bias in AI is a part of a larger effort, in which NIST seeks to engage private and public sector organizations and individuals in discussions about building blocks for trustworthy AI systems. The focus in this work is to identify the associated measurements, methods, standards, and tools necessary to implement those building blocks when developing, using, and testing AI systems.

Information about NIST's other AI efforts is available here: https://www.nist.gov/topics/artificial-intelligence

August 18 Workshop: What to expect

Introduction

The intent of this workshop is to provide an inaugural venue for discussions about what constitutes the key building blocks for trustworthy AI systems. Bias remains a key but still insufficiently defined building block. We hope to enable panel and group discussions about bias, in a manner that will provide the necessary insights to move the AI community closer to agreement on its definition. This workshop will also help to build a collaborative environment for NIST's multi-faceted work in the broader arena of Trustworthy AI, which includes foundational and use-inspired research, evaluation, standards, and policy engagement.

The one-day workshop will consist of two panel discussions and two smaller-group breakout sessions. These discussions are an opportunity for diverse participant groups to be heard, drive debate, and build a shared understanding. The topic areas for the day are centered around data and algorithmic bias. There is limited agreement and understanding on whether these two topics overlap, or how they converge and interact. We hope that at the end of the workshop we will gain better insight into those foundational questions.

Panel discussions

To engage the community in foundational conversations and develop a shared understanding of bias in AI, the NIST Workshop on Bias in AI will have two panel discussions throughout the day.

- The morning panel--***Juggernaut: Addressing data bias challenges in AI***--will focus on the key challenges of dealing with the bias inherent in the datasets used in AI, including where, how and when bias starts to play a role.
    - The morning panel will be moderated by Darrell West. Panelists are:
        - Andrew Burt
        - Alexandra Chouldechova
        - Fernando Diaz
        - Teresa Tung
    - The grounding for this discussion centers around the following questions:
        - If data is the fuel of the algorithm and is reflective of society (and its biases), can we remove them in the AI development process?
        - How can we ensure we aren't inserting additional biases into the process?
    - The panel will engage in a deeper discussion about:
        - How is bias exhibited in data?
        - What is needed to measure bias in data?
        - How do we decide what to tackle first?
        - How can issues of data access and availability affect AI bias?
        - What does success look like and how can we best track our progress?
        - What are you most concerned about and what are the biggest barriers to success at this point?
- The afternoon panel--***Algorithmic bias is in the question, not the answer: Measuring and managing bias beyond data***--will focus on how bias can affect modeling and algorithmic decisions and outcomes.
    - This panel will be moderated by Joshua Kroll. Panelists are:
        - Aylin Caliskan
        - Abigail Jacobs
        - Nicol Turner Lee
        - Kush Varshney
    - The grounding for this discussion centers around the following questions:

        - Algorithms are highly dependent on data, which can be biased. But how does the algorithmic process insert additional biases?
        - Is it possible to use algorithms to mitigate data biases?
    - The panel will engage in an in-depth discussion about algorithmic bias, including:
        - How can measurement help us understand algorithmic bias and its origins?
        - Is debiasing a suspect approach?
        - How can we use tools and standards to measure bias in algorithms?
        - What is the role of problem specification in the AI lifecycle?
        - What are the dominant concerns and barriers to measure and manage algorithmic bias?

Breakout sessions

- Following each panel discussion, participants can join their assigned breakout session groups to dig deeper into the conversation about bias in AI, and build on the topics raised during the panel sessions. All participants will be assigned to one of five breakout sessions, each of which will focus on the same key questions. Each session will be hosted by a designated facilitator and scribe.
- The five morning breakout sessions will focus on data bias and build on topics from the panel session. Participants will have an opportunity to provide their insights in a facilitated discussion about the following questions:
    - What is the "right data"?
    - What are the biggest barriers to success at this point?
    - How can technology developers and practitioners effectively work together and inform each other to mitigate data bias?
- The five afternoon breakout sessions will focus on bias in algorithmic modeling and build on topics from that panel session. Participants will have an opportunity to provide their insights in a facilitated discussion about the following questions:
    - Algorithms are highly dependent on data, which can be biased. But how does the algorithmic process insert additional biases?
    - Is it possible to use algorithms to mitigate data biases?

After each breakout session there will be a group report-out by the facilitator in the plenary room.

Registrant response themes

During the registration process, all workshop registrants were asked to provide their own definition of bias in AI. An evaluation of the responses indicated that, in general, attendees have a broad sense of the importance of addressing bias in AI--the *why*--with less consistent definitions of *what* bias in AI is, or *how* to measure it. The following themes emerged:

- There is a general recognition that bias exists in both society and in AI systems. Respondents acknowledged general issues with

    - the availability of data,
    - system output or results,
    - human oversight of AI systems.
- Respondents highlighted the importance of attending to cues that an algorithm is biased, and understanding the sources and indicators of AI bias. Sub-topics include training data challenges, and unintended results when AI is implemented in the real world.
- There is general concern about how models and algorithms echo or mirror societal biases. Respondents noted the potential for models to amplify existing biases within society, especially racial and gender biases.
- There is a general sense that the "right" input and quality metrics are necessary. Respondents often used the phrase, "garbage in, garbage out," when describing the importance of high-quality training data.

Bias in AI: Related activities

Achieving stakeholder consensus around the organizing principles and key terminologies used when discussing, developing, and implementing AI is a necessary foundation for standards development in trustworthy AI. With this goal in mind, NIST intends to develop a report that will focus on a taxonomy of concepts and terminology in AI Bias. The taxonomy, built on and integrating previous work in AI bias, will be arranged in a conceptual hierarchy that includes key factors associated with the lifecycle of AI applications. Taken together, the terminology and taxonomy are intended to inform future standards and best practices for mitigating bias in AI applications, and to establish a common language and understanding in this area.

Participants are encouraged to share insights and recommendations regarding relevant taxonomies and terminologies in bias in AI on the day of the workshop.

This report will serve as a common reference point for future activities designed to understand the building blocks of bias in AI, and developing Trustworthy AI. To learn more please be on the lookout for more information via email and on the NIST Bias in AI webpage https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-free-bias

Literature survey of Bias in AI

NIST has been conducting a literature review of current and relevant articles on the topic of bias in AI to inform the workshop agenda and discussion topics. To-date, over 240 articles have been reviewed. This bibliography will be shared in a literature survey report by the end of 2020.

NIST reviewed materials from frequently-cited, shared, and cross-referenced pieces focusing on bias within technologies that use artificial intelligence. This review incorporated content that described AI bias from societal contexts, pre-existing technologies, development processes, and other factors that

influence AI development, implementation, and/or adaptation. To ensure a cross-section of perspectives, literatures are reviewed 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, across current and future AI implementations. Topics included in the current literature survey relate to understanding the human- and systems-level roles in identifying, understanding the cause of, and mitigating or preventing bias in AI; and understanding how societal biases affect past, current, and future technologies.

Preliminary descriptives of NIST AI Bias literature review	
Article Type	
Academic Journal Article	129
Books or edited volumes	5
Conference paper or presentation	39
Hearing or letter	5
Magazine article, News article, or Web page	33
Report	32
Total Articles	243

In an attempt to understand the larger expanse of this topic area, we visualized the collaborative connections between communities of different domains who publish about AI bias. This was achieved by searching articles that have “AI Fairness” or “AI Bias” in their title or abstract on Microsoft’s academic research API. This initial search returned 1000 documents from the API, which were then analyzed by co-occurrence of fields of study (keywords) in VOSViewer 1.6 (van Eck & Waltman 2010). Of the 4020 keywords extracted, the top 50 frequencies were selected for visualization. The top 20 of these keywords are listed in the table below along with their weighted degree centrality (a measure of how many times they were used and co-occurred with the other top 50 keywords). (e.g. If a keyword occurred twice with five other keywords on the list it would have a weighted degree centrality of ten).

Keyword	No. of Occurrences in Corpus	Weighted Degree Centrality
Computer Science	357	630

Artificial Intelligence	254	613
Medicine	116	189
Psychology	102	129
Machine Learning	73	229
Humanities	60	47
Transparency	57	163
Biology	56	43
Internal Medicine	46	114
Deep Learning	45	160
AI Systems	45	109
Population	45	77
Sociology	43	59
Accountability	42	139
Materials Science	41	17
Data Science	39	99
Political Science	39	49
Mathematics	39	34
Business	38	60
Artificial Neural Networks	36	99

The figure below is a network visualization of the linkages in the corpus of documents on AI fairness or AI bias which produces a scientific 'map' of the subfields who are discussing these topics the most. Note that only 49 of the original 50 nodes are included in this network as Environmental Science, which had the least amount of occurrences at 23, had no network connections to the other keywords and was removed from the visual. Node color represents the years the keywords were most mentioned in the corpus and follow the legend in the bottom right. Node size represents the total number of occurrences and the size of the linkages shows the amount the keywords were used together.

Besides showing 'clusters' of scientific sub-fields discussing the issues of AI fairness & biases, it is apparent that moving from right to left in the visualization you see newer subfields beginning to achieve importance in the academic discourse surrounding AI fairness and biases. Also, these fields often overlap with social science disciplines such as Sociology and Business which are less integrated with the medical and physics literature. Insights such as these from the Network Science literature will be used to both inform and provide quantitative analysis of the forthcoming NIST report. Such analysis of the relevant literature will inform cross-discipline standards that speak to both nascent and more mature discussions around addressing bias in artificial intelligence.

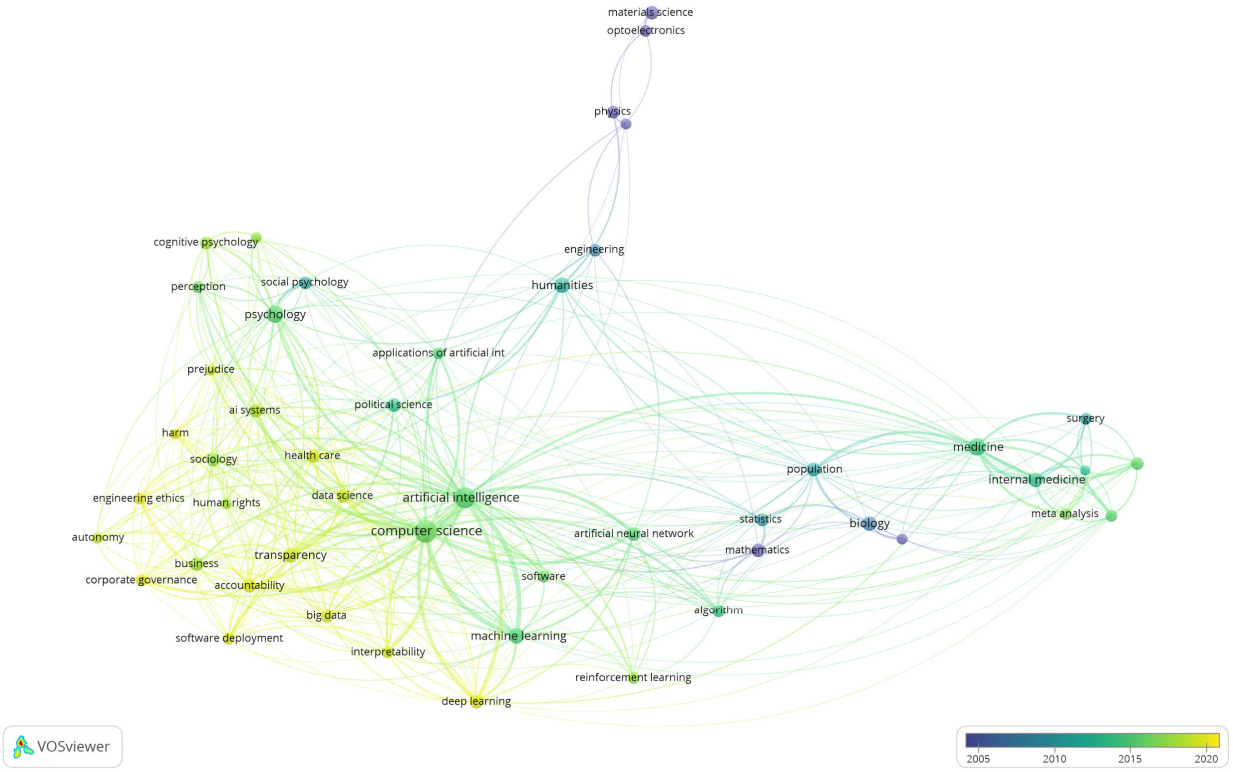

Figure 1. Network visualization of linkages in a corpus of documents on AI fairness or AI bias.

AI lifecycle

The AI lifecycle -- the cyclical process through which AI products move across phases of development -- is foundational in understanding the ways bias can affect AI. For example, it can allow us to investigate the phases of the AI lifecycle that might be most susceptible to bias. However, **there is no global or industrial standard for the AI lifecycle.** There are currently a variety of AI lifecycles in use across multiple sectors and regions. In order to reach a shared understanding of bias in AI--and thus support future work in developing standards around it--the community must operate from a shared understanding of the AI lifecycle.

There are several versions of the AI lifecycle to which industry stakeholders often refer. The most frequently cited versions of the AI lifecycle across literatures include models developed by the Centers of Excellence (CoE) at the US General Services Administration[1] and the Organisation for Economic Co-operation and Development (OECD)[2]. It is worth noting that another model of the AI lifecycle is currently under development with the Joint Technical Committee of the International Organization for Standardization and the International Electrotechnical Commission (SC 42).

These lifecycle models range from three phases to eight, with various degrees of specificity and interaction depicted within and across each phase. Links to source materials and diagrams of the high-level phases within each of the two most oft-cited lifecycle models are included below.

The community conversations at the NIST Workshop on Bias in AI will allow us to build on both the shared and disparate understandings of the AI lifecycle. We invite participants to explore the discrepancies between these models and identify opportunities for consensus. Source information and additional details regarding each AI lifecycle model can be found on the NIST AI Bias webpage at https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-free-bias

There are several commonalities between these AI lifecycle models, including:

- **The AI lifecycle begins in early design stages, before algorithms are involved.** Across lifecycle models, this initial stage in the AI lifecycle includes planning, pre-design, identification of data, problem specification, and background research.
- **Validation plays a role.** Although the AI lifecycle versions incorporate slightly different terminology and placement within the lifecycle, each specifies the need for validation to be built into the lifecycle, but exactly where and how remains an open question.
- **Deployment is intentional and involves more than pressing "go."** Each version of the AI lifecycle model includes a deployment-specific stage, which often entails user engagement, training, and informing stakeholders about updated product roll-outs.
- **Ongoing monitoring and evaluation is essential.** Most of the AI lifecycles clearly identify one or more phases of monitoring, which underscores the importance of evaluation planning at project outset. It is currently unclear how these monitoring steps could most effectively impact the identification of bias or necessitate steps that should be taken to mitigate it.

Notable differences between the AI lifecycle models include:

- **Specificity of lifecycle phases.** Some AI lifecycles use more general terms to encompass a variety of tasks, while others break down the phases into more nuanced steps.
- **Use of colloquial terminology.** Terms are used differently across lifecycles. For example, the term "development" is used to describe a variety of activities at different points in the AI lifecycle from model to model.

---

[1] Centers of Excellence at the US General Services Administration. (n.d.). *GSA.* Retrieved from https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf.

[2] OECD. (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. *OECD.* Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

- **Sequencing of events.** While the most general sections of each AI lifecycle align (e.g., design tends to happen early on, while monitoring happens toward the latter phases), more specific tasks/phases--such as validation, deployment, and evaluation--are placed at different stages from lifecycle to lifecycle.

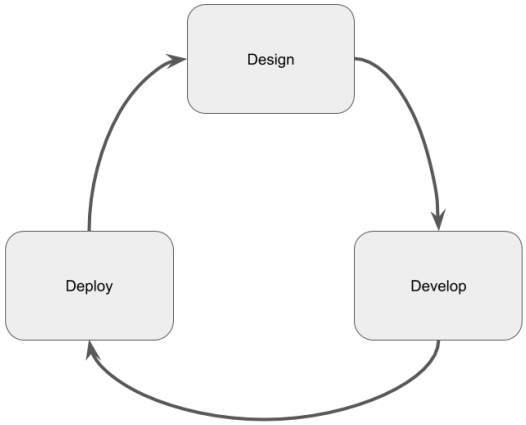

Figure 2. CoE AI Lifecycle Model.
Source: Centers of Excellence (CoE) at the US General Services Administration. (n.d.). *GSA.* Retrieved from https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf.

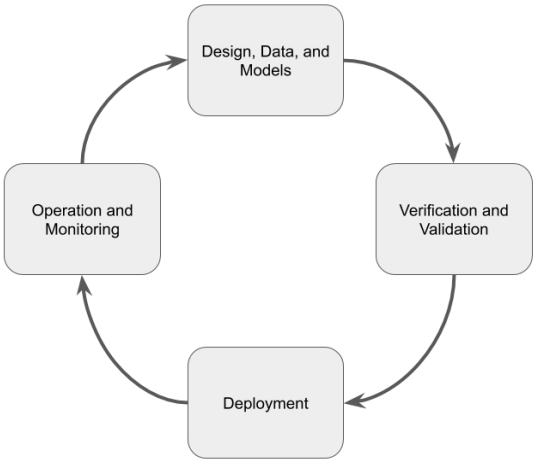

Figure 3. OECD AI Lifecycle Model.
Source: OECD. (2019). OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. *OECD.* Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

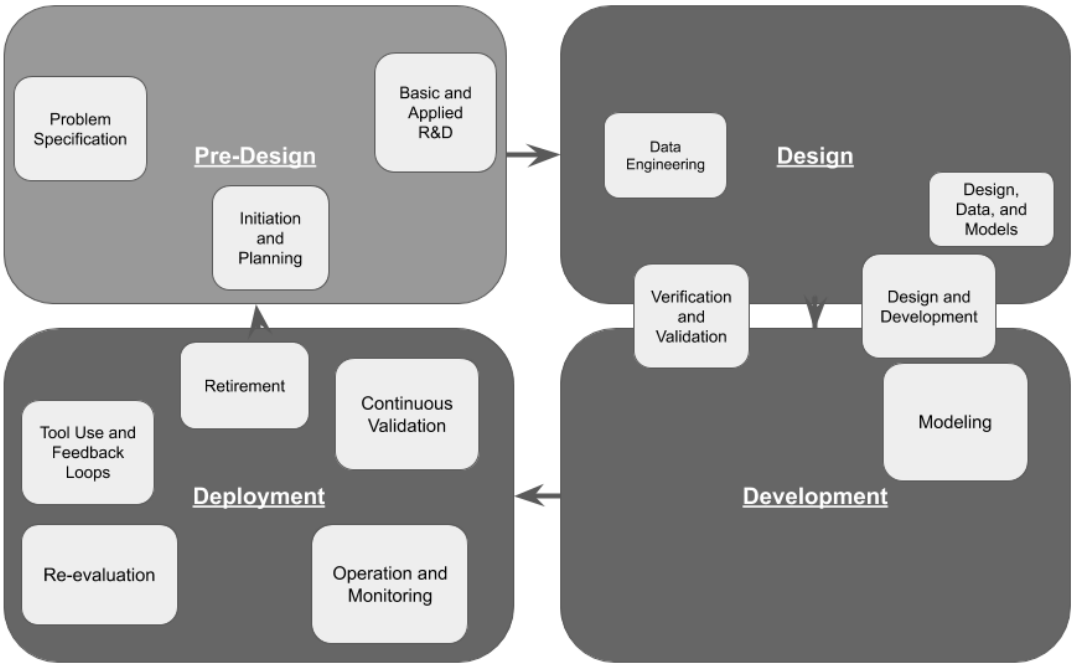

Figure 4. Current approaches to the AI lifecycle: Including areas of overlap across CoE and OECD models.
Sources:
(1) Centers of Excellence (CoE) at the US General Services Administration. (n.d.). *GSA.* Retrieved from https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf.
(2) OECD. (2019). OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. *OECD.* Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Metrics and measurements

To adequately understand, assess, and improve industry standards in addressing bias in AI, developers and practitioners across sectors must be able to reliably gauge its occurrence. However, there is currently no cross-disciplinary or cross-sector consensus in approaches to identifying or validating measurements, metrics, and key indicators of bias, or how social data should be measured or understood in context. This workshop aims to address this gap through identifying the associated measurements, methods, standards, and tools necessary to prevent, identify, and mitigate bias in AI systems.