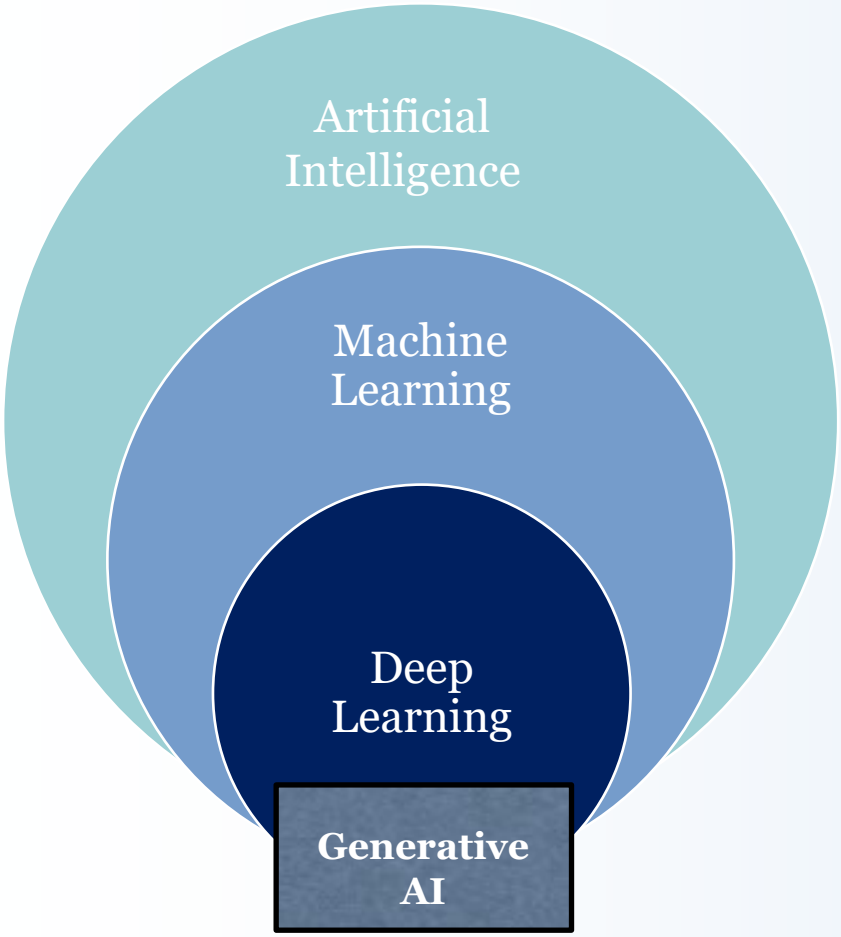
BROOKINGS | QUALITY. INDEPENDENCE. IMPACT.

The AI Equity Lab: Identifying and Mitigating Online Biases

Dr. Nicol Turner Lee
Senior Fellow, Governance Studies
and Director, Center for Technology Innovation
The Brookings Institution

December 4, 2023

AI: The New Ecology



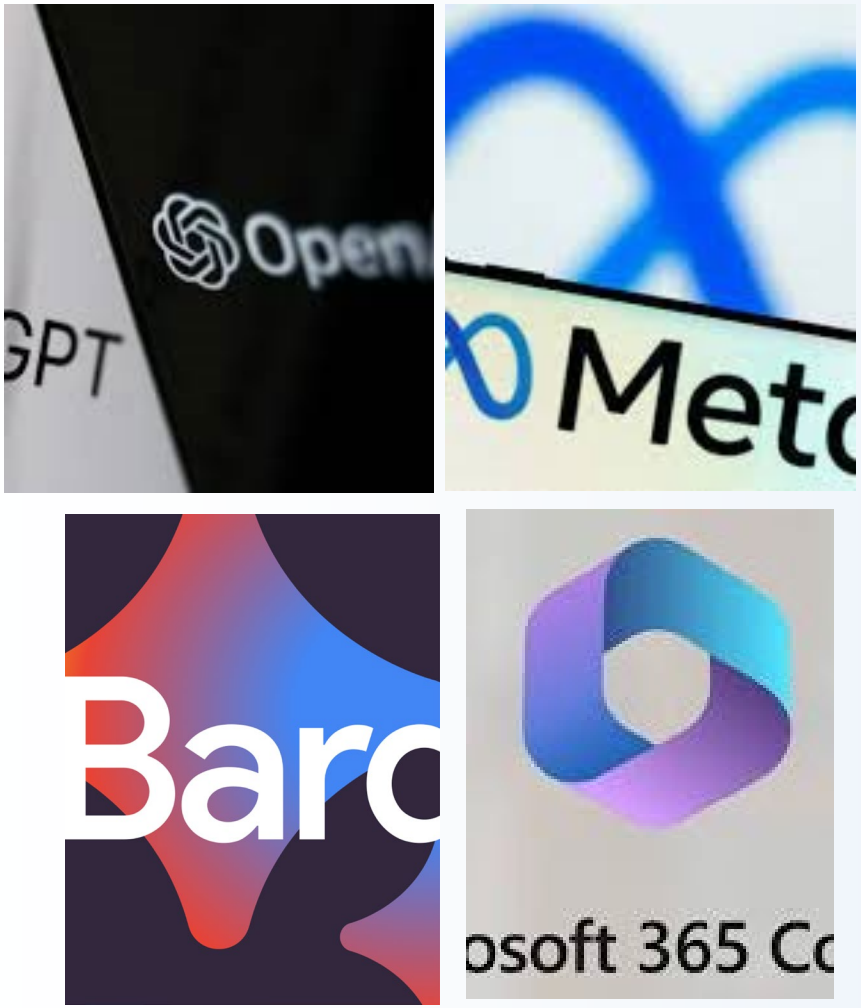
Artificial Intelligence (AI) systems powered by algorithms – sets of instructions to perform tasks

Machine Learning (ML) – the automation of repetitive tasks and functions

Deep Learning – A subset of ML that permits software to train itself to perform tasks, e.g., speech and image and facial recognition

Generative AI – Trained on Large Language Models (LLM) that permits predictive and summative insights from larger amounts of data.

Taken together, these emerging technologies can efficiently and quickly solve a host of existing and unforeseen scientific, social, and technical problems.

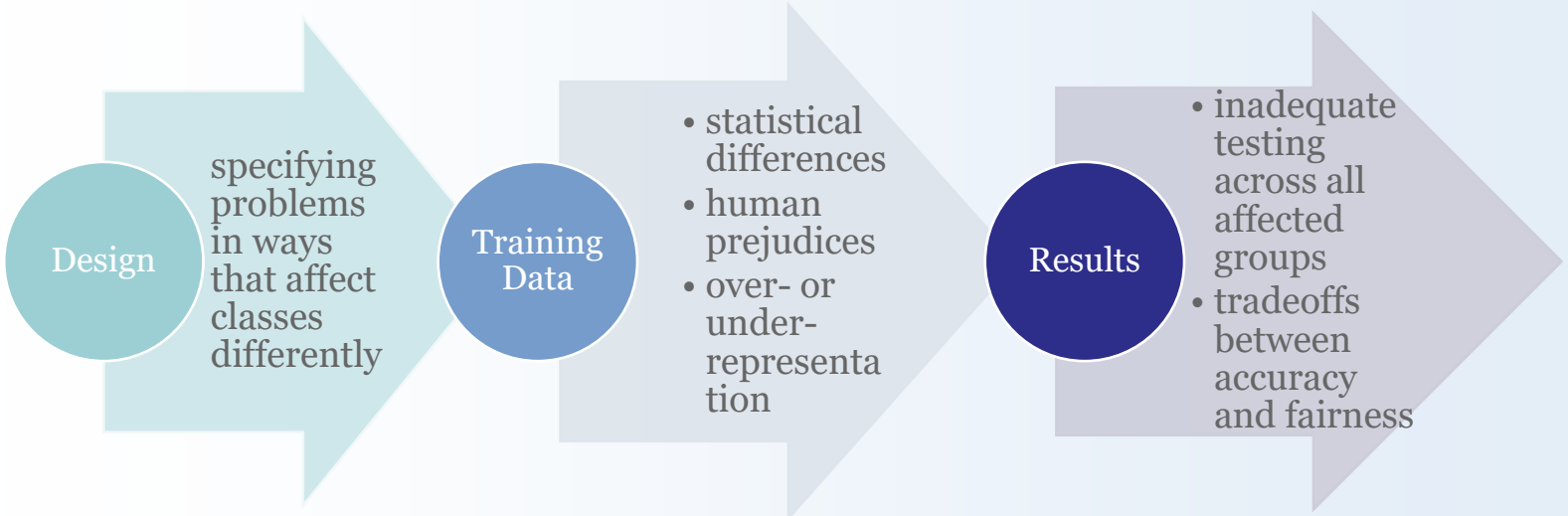


Generative AI is beginning to surface these and other questions

- "Eyes in the back of the head" when it comes to scraping and analyzing hyper-textual data, curated images and voices.
- Functions of handling, sorting, and analyzing Large Language Models (LLMs).
- Technology powers chat boxes, predictive generative content, and other applications have the potential for bias, copyright infringement, misinformation, and reduced intellectual integrity.

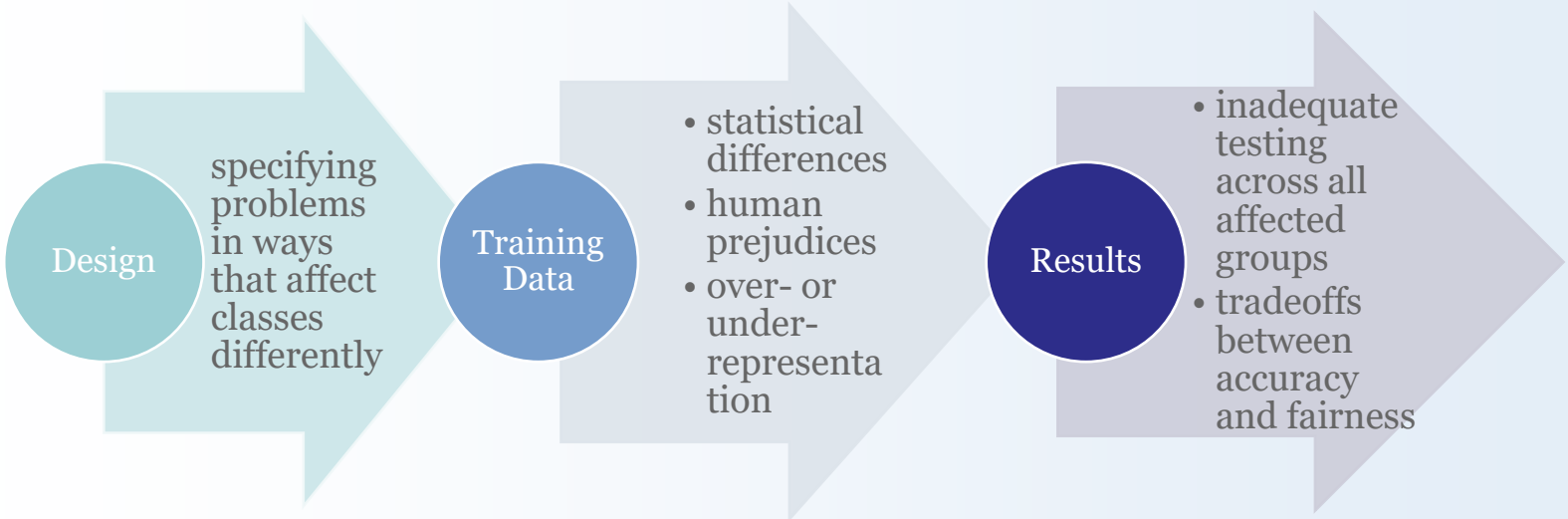
Bias in Algorithms

Bias can be introduced at any stage of the life cycle of the algorithm



Bias in Algorithms

Bias can be introduced at any stage of the life cycle of the algorithm



Inferences from data about people – including their identities, demographic attributes, preferences, and likely future behaviors – can have adverse impacts on disfavored groups beyond just protected classes.

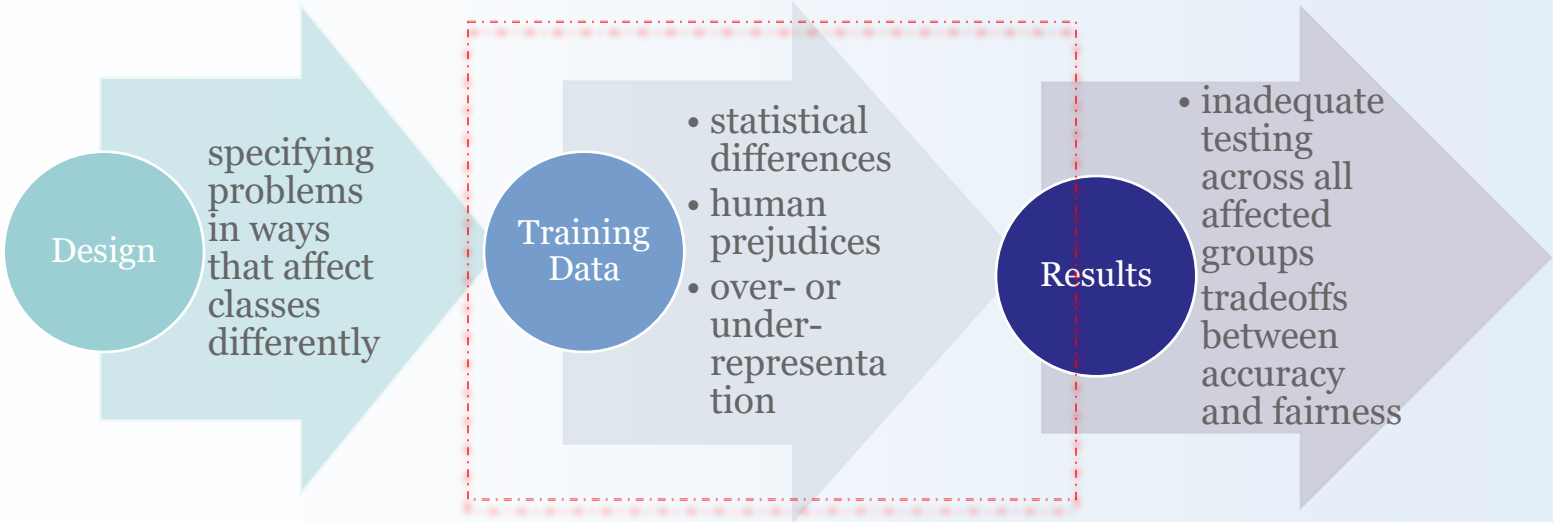
Source: Nicol Turner Lee, Paul Resnick, and Genie Barton, May 22, 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce harm. Brookings, Center for Technology Innovation, Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms | Brookings.

Inferential modeling and assessments

- Inferences from data about people, including their identities, their demographic attributes, their preferences, and their likely future behaviors can have an adverse impact on any disfavored groups
- Point of view discrimination and targeting is beyond current scope but is also important
- Source: Kearns, Michael and Roth, Aaron, 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press)

Source: Nicol Turner Lee, Paul Resnick, and Genie Barton, May 22, 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce harm. Brookings, Center for Technology Innovation, Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms | Brookings.

Data Quality and Trauma



The data we decide to use for training the "inputs and outputs" of AI models are explicitly and implicitly biased and pre-determined.

They are also not immune to historical and systemic inequalities.*

*Source: Renee Cummings, University of Virginia

Where biases show up?

Ads

Where biases show up?

Where biases show up?

- Ads
- Hiring and employment

Where biases show up?

- Ads
- Hiring and employment
- Health care

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications
- Policing

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications
- Policing
- Search engines

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications
- Policing
- Search engines
- Financial services

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications
- Policing
- Search engines
- Financial services
- Housing

Where biases show up?

- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications
- Policing
- Search engines
- Financial services
- Housing
- Lifestyle

Where biases show up?

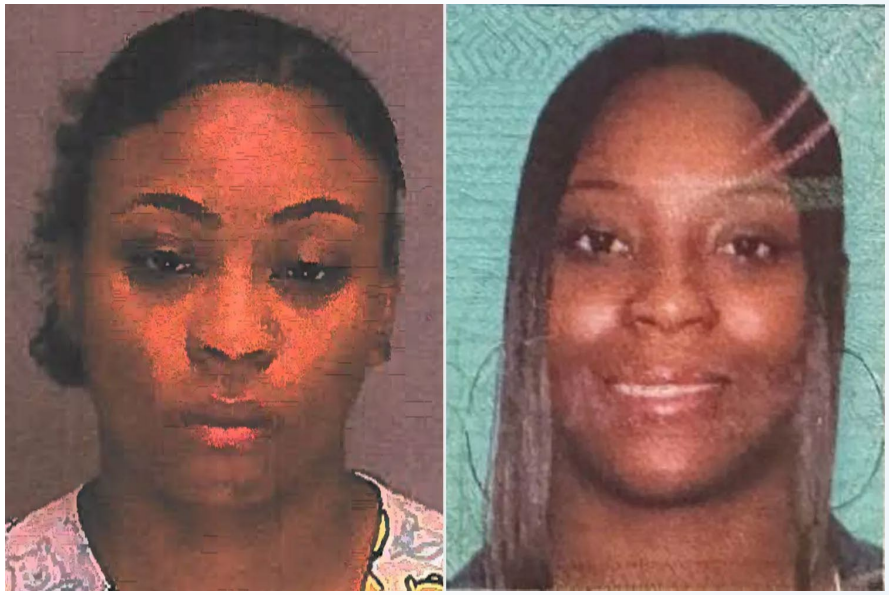
- Ads
- Hiring and employment
- Health care
- Education
- College admissions and testing
- Criminal justice applications
- Policing
- Search engines
- Financial services
- Housing
- Lifestyle
- And more….

Fairness is an Ethical & "Everybody" Question

- Fairness, overall, is an **elusive concept** to define.
- Developers, deployers, and users of AI can try to achieve greatest degree of accuracy and fairness across various measures, but **tradeoffs** will still happen.
- Where an algorithm could have a negative effect on human beings, decisions must occur within an **ethical framework** that reflects societal expectations, respects individual privacy, and minimizes harm to subjects and communities.
- Determining what **tradeoffs** are most "fair" can be subjective because human beings are flawed and driven by their own norms, values, and assumptions.
- How we interrogate these models in specific use cases with varying degrees of **risk** matters because it reduces consequential tradeoffs to human subjects and their communities.

Use Cases Gone Awry

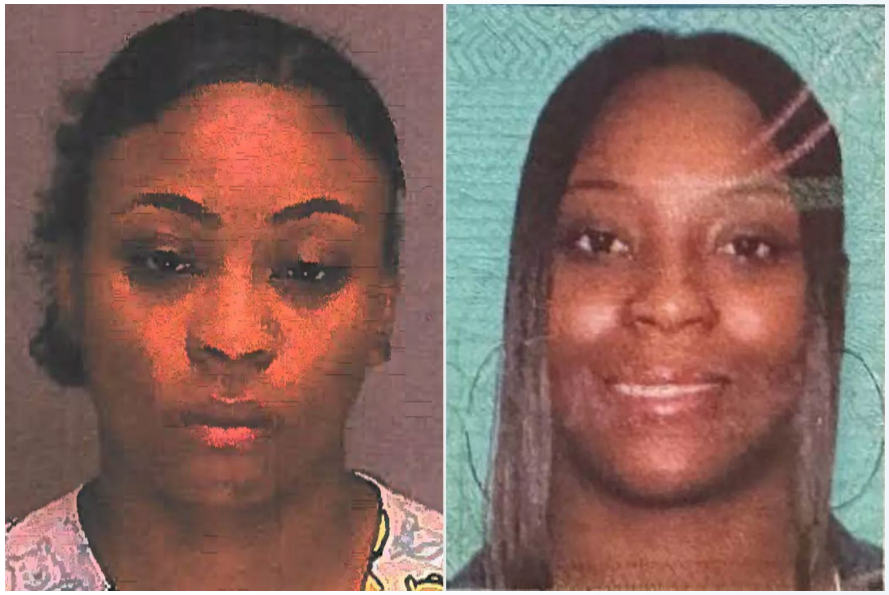
- Facial recognition technology
- Pulse oximeters
- Employment and hiring



2015 photo of Porcha Woodruff led to her 2023 arrest

The Detroit Police Department used facial recognition technology to identify Porcha Woodruff as a car jacking suspect.*

*Source: Kelly Kasulis Cho, August 7, 2023, "Women Sues Detroit after facial recognition mistakes her for crime suspect." The Washington Post.



2015 photo of Porcha Woodruff led to her 2023 arrest

The Detroit Police Department used facial recognition technology to identify Porcha Woodruff as a car jacking suspect.*

She was arrested in front of her children and is now suing the city of Detroit.

The Woodruff family

*Source: Kelly Kasulis Cho, August 7, 2023, "Women Sues Detroit after facial recognition mistakes her for crime suspect." The Washington Post.

By delivering a pulse of light through the skin, the pulse oximeter tracks blood oxygen levels.

By delivering a pulse of light through the skin, the pulse oximeter tracks blood oxygen levels.

But these devices do not work well on darker skinned populations due to the melanin in the skin, which interferes with the accurate detection of oxygen, and other critical conditions.*

Source: The Federation of American Scientists - An Overdue Fix: Racial Bias and Pulse Oximeters - Federation of American Scientists (fas.org)

Various hiring tasks are being enabled by AI, including pre-screening, skills testing, and interviewing.

Various hiring tasks are being enabled by AI, including pre-screening, skills testing, and interviewing.

- **Black professionals receive 30% to 50% less job call backs when their resumes contain racial or ethnic identity.***
- **Black and Hispanic applicants are more likely to be screened out by video interviews due to their facial expressions (e.g., eye contact, smiling, etc.).**
- **Black and Hispanic workers are also most likely to be replaced by automation.***

*Source: Dawn Zapata, June 18, 2021. New study finds AI-enabled anti-Black bias in recruiting. Harvard Business Review, New study finds AI-enabled anti-Black bias in recruiting - Thomson Reuters Institute

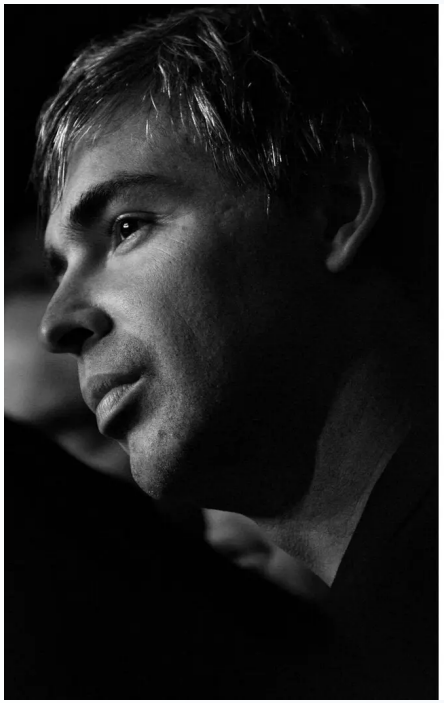
Broad risks and concerns are socio-technical in the U.S.

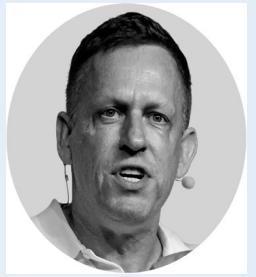
- **Societal**
  - **Biased decision-making:** AI-driven algorithms often contain gender, racial, or other implicit/inferential biases that reinforce systemic discrimination.
  - **Deliberate misuse:** malicious actors may spread disinformation, create deepfakes, and conduct unauthorized surveillance or profiling.
  - **And more**: job displacement, data privacy, plagiarism/copyright/IP, carbon emissions & environmental impact of training large models.
- **Technical**
  - **Nonrepresentative and traumatized datasets**: Without diverse representation, the data can be an insufficient, leading to differential treatment and potentially disparate impact.
  - **Tradeoffs of efficiency**: The productivity outcomes are often seen as surpassing the intended and unintended consequences on subjects.
  - **Minimal standards**: The lack of benchmarked performance metrics make it harder to interpret and assess fair, ethical, and lawful autonomous systems.

AI systems are also geo-politically rooted

- **The need for more globally inclusive dialogues**
  - **Digital sovereignty**: Who sits at the table will determine who has control over the AI supply chain, from data to hardware to software.
  - **Colonial repression and exclusion**: Issue of national security, interoperable systems, and the push for AI for new markets invasion rather than participation/production.
  - **And more**: climate and emissions, data workers, and digital rights for culturally distinguishable continents, regions, and countries.

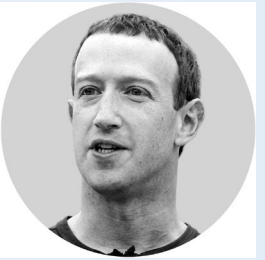
NYT: "Who's Who Behind the Dawn of the Modern Artificial Intelligence Movement." (12/3/23)





AI dialogues must be inclusive - Women of color in AI

The Brookings AI Equity Lab

- **A project that:**
  - » **convenes diverse stakeholders in conversations that result in pragmatically possible ideas for more equitable AI development and deployment;**

The Brookings AI Equity Lab

- **A project that:**
  - **convenes diverse stakeholders in conversations that result in pragmatically possible ideas for more equitable AI development and deployment;**
  - **creates a responsible and inclusive ecosystem for current and future dialogues on AI's public interest impacts;**

The Brookings AI Equity Lab

- **A project that:**
  - **convenes diverse stakeholders in conversations that result in pragmatically possible ideas for more equitable AI development and deployment;**
  - **creates a responsible and inclusive ecosystem for current and future dialogues on AI's public interest impacts;**
  - **develops a resource for best-in-class research and public policy proposals at the domestic and global levels that advance inclusive conversations on AI technical cadence, regulation, and programs in both the private and public sectors;**

The Brookings AI Equity Lab

- **provides a platform for exploring research and policy deliverables via public events and off-the-record convenings, e.g., roundtables, to establish an echo chamber for future work focused on non-discriminatory, anti-racist, and anti-colonial AI services, products, and ecosystems; and,**

The Brookings AI Equity Lab

- **provides a platform for exploring research and policy deliverables via public events and off-the-record convenings, e.g., roundtables, to establish an echo chamber for future work focused on non-discriminatory, anti-racist, and anti-colonial AI services, products, and ecosystems; and,**
- **develops a "Hidden Figures" repository that gathers, shares, and amplifies the perspectives of leaders under-represented and less visible in AI research and policy debates. (*Winter 2024*)**

Phase One – Workshopping critical areas to advance nondiscrimination

- **Interdisciplinary and cross sector workshops on critical services with socio-economic and social mobility outcomes with anticipated guidance on redressing consumer harms and implicating systemic inequalities for consumers.**
  - Health care
  - Employment
  - Education
  - Housing
  - Criminal justice
  - Voting and election security

Phase One – Workshopping critical areas to advance nondiscrimination

- **Deliverables**
  - Short policy and guidance briefs that address:
    - The problem
    - Opportunities and challenges
    - Risks/harms – individual and collective
    - Existing safeguards/guardrails
    - New and more adaptable policy and programmatic proposals, i.e., what can be done to effectuate change.

Phase One – Workshopping critical areas to advance nondiscrimination

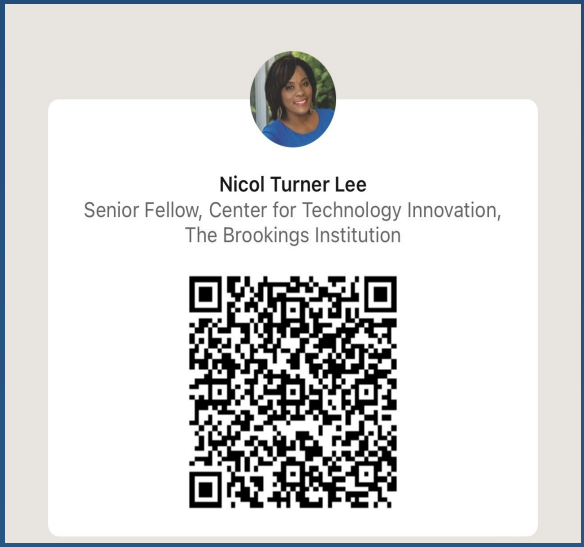
- **Deliverables**
  - Short policy and guidance briefs that address:
    - The problem
    - Opportunities and challenges
    - Risks/harms – individual and collective
    - Existing safeguards/guardrails
    - New and more adaptable policy and programmatic proposals, i.e., what can be done to effectuate change.
  - Comprehensive Brookings report with themes from guidance documents, and broader research and policy implications. (*Spring 2024*)

BROOKINGS | QUALITY. INDEPENDENCE. IMPACT.

Find me: Dr. Nicol Turner Lee
@drturnerlee (Twitter)
nturnerlee@brookings.edu (email)

Turner Lee, Nicol. Digitally Invisible: How the internet is creating the new underclass (Brookings Press, 2024).

Follow me on Linkedin to get publication updates.



BROOKINGS | QUALITY. INDEPENDENCE. IMPACT.

In-person | December 4, 2023, 2:00 - 3:30 pm EST | #EquitableAI

BLACK WOMEN IN AI: BUILDING A MORE INCLUSIVE AND EQUITABLE FUTURE

Presentation and moderator

NICOL TURNER-LEE
Senior Fellow and Director,
Center for Technology Innovation,
The Brookings Institution

Panel discussion

RENÉE CUMMINGS
Nonresident Senior Fellow,
Center for Technology Innovation;
Professor, Practice in Data Science,
University of Virginia (UVA)

FAY COBB PAYTON
Visiting Scholar and
Special Advisor on
Inclusive Innovation,
Rutgers University

MUTALE NKONDE
CEO,
AI for the People;
Visiting Policy Fellow,
Oxford Internet Institute

BROOKINGS