WOMEN WHO
CODE/connect

Mitigating Biased Algorithms in the Real World

Hinda Haned, PhD.

Senior Lead Data Scientist
Owls & Arrows | University of Amsterdam

May 2022

Civic AI Lab

An engaging Society where all citizens have equal opportunity to participate and engage in a fair, transparent manner

UNIVERSITY OF AMSTERDAM

CAIL

WOMEN WHO CODE

Algorithmic fairness: why bother?

Controversial AI applications

BBC Sign in News Sport Reel Worklife Travel Future M

NEWS

Home Video World UK Business Tech Science Stories Entertainment & Arts

Amazon scrapped 'sexist AI' tool

10 October 2018 Share

2,780 views | Jan 28, 2019, 07:20pm Forbes

Amazon Refuses To Quit Selling 'Flawed' And 'Racially Biased' Facial Recognition

Zak Doffman Contributor

NOS Nieuws Sport Uitzendingen

VN-rapporteur zeer bezorgd over Nederlands opsporingssysteem voor uitkeringsfraude

De VN-rapporteur voor de mensenrechten Philip Alston heeft ernstige zorgen over Nederland. De reden is een systeem dat uitkeringsfraude moet opsporen. In een brief aan de rechtbank in Den Haag schrijft Alston dat het systeem in strijd is met de mensenrechten omdat het mensen met weinig geld en mensen met een migratie-achtergrond discrimineert.

Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

by Karen Hao Apr 5, 2019

Compliance

Compliance with General Data Protection Regulation (GDPR – 2018) – article 22

*The data subject shall have the **right not to be subject to a decision based solely on automated processing**, including **profiling**, which produces legal effects concerning him or her or similarly significantly affects him or her*

Compliance

Date	Organisation	Amount	Issued by	Reason(s)
2021-12-16	Psykoterapiakeskus Vastaamo	€608,000	Finland	Failure to protect sensitive medical data.[74]
2021-09-02	WhatsApp Ireland Ltd	€225 M	Ireland	[73]
2021-06-16	Amazon Europe Core Sarl	€746,000,000	Luxembourg (CNPD)	The largest fine for violating GDPR to date.[71][72]
2021-05-11	PVV (Overijssel)	€7,500	The Netherlands (AP)	Violation of Articles 4(12), 9(1) GDPR and 33(1) GDPR by unauthorised disclosure of a mailing list containing 101 email addresses, and failing to notify this breach to the DPA. The email addresses constituted special category data revealing political party opinions.[69][70]
2021-05-06	Owner's association in Iasi	€500 (RON 2,463.30)	Romania (ANSPDCP)	Violation of Articles 58(1)(a), 58(1)(e), 83(5)(e) GDPR as well as of Article 8 of Government Ordinance No 2/2001, by violating the obligation to cooperate with the DPA during an investigation by failing to provide the information requested[67][68]

https://en.wikipedia.org/wiki/GDPR_fines_and_notices

Complex systems raise concerns

- Why this ad?
- Why this discount?
- Why this recommendation?

- Why was I rejected?
- Can I change the outcome?
- When will the system fail?

Algorithmic fairness: definitions

What is bias

- Systematic errors that create unfair outcomes
- Sources: algorithm design, biased data collection or selection
- Algorithms learn and perpetuate bias

What is fairness?

- Fairness is concerned with how outcomes are assigned to particular group of individuals
- Core principle: avoid bias even if it is supported by data, as to avoid the perpetuation of existing discrimination
- Fairness is a political construct: someone decides to avoid (direct or indirect) harms

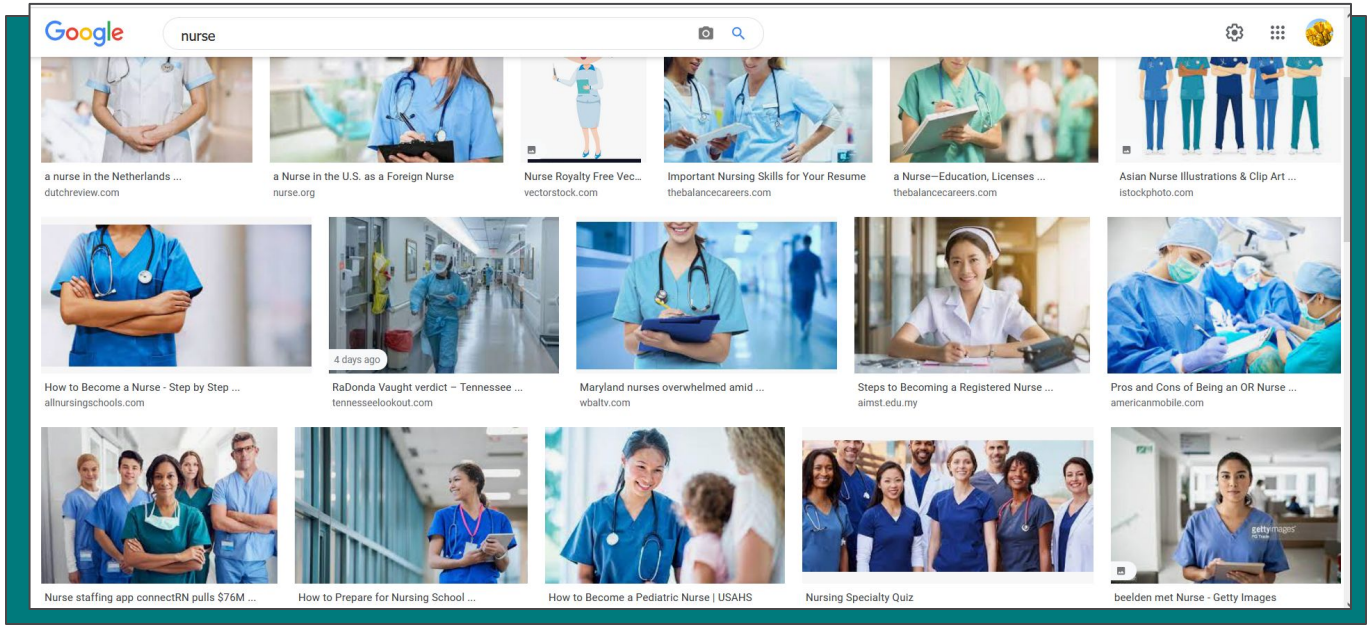
Types of harm

- **Harm of allocation**: when a system allocates or withholds certain groups, an opportunity or resource. Economically oriented view (e.g. who gets a discount, who gets hired),
- **Harm of representation**: systems reinforce the subordination of certain groups along the lines of identity like race, class, gender etc. (e.g. search results biased against a group),
- **Reduced quality of service:** when a system produces degraded user/customer experience based on their personal characteristics (e.g stereotyping and denigration)

Adapted from: Kate Crawford's NIPS 2017 Keynote presentation: the trouble with Bias & Challenges of incorporating algorithmic `fairness' into practice: FAccT 2019 tutorial

Historical Bias

Representation Bias



Queried APR22

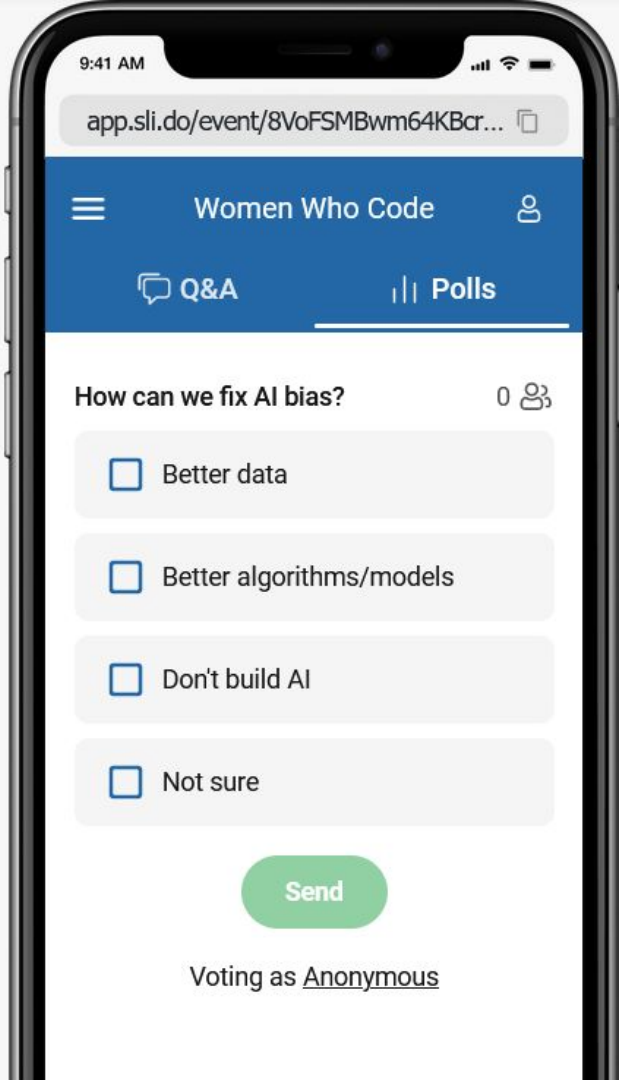
Measurement bias

How can we avoid bias?

Question time

slido.com

#1415235



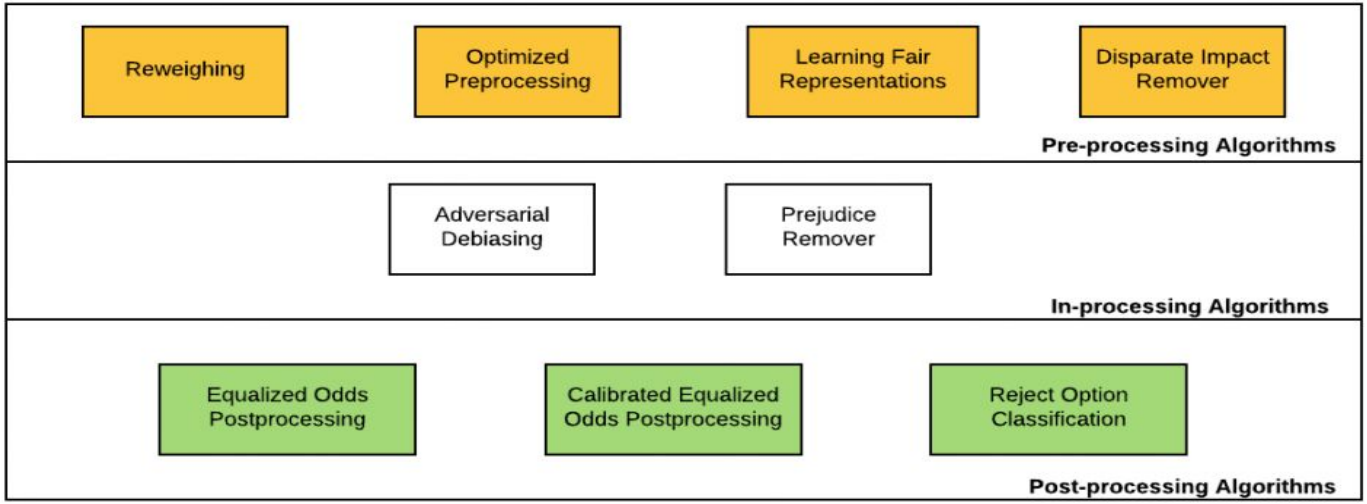
Ethics

Ethics Guidelines for Trustworthy Artificial Intelligence - European Commission

FACT-AI

https://facctconference.org/

Mitigation algorithms: *remove or reduce bias in data and model outputs*

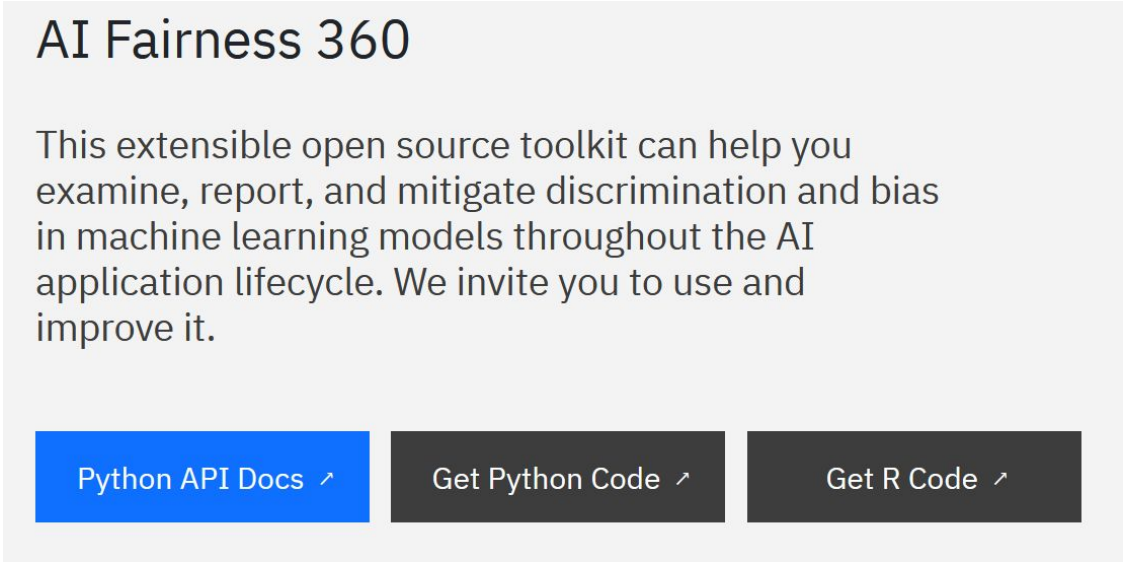


Bias Mitigation Strategies for ML Models

Source: https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies

Mitigation algorithms: *remove or reduce bias in data and model outputs*

- Known sensitive attributes
- Defined fairness objectives
- Supervised learning tasks
- Binary classification problems

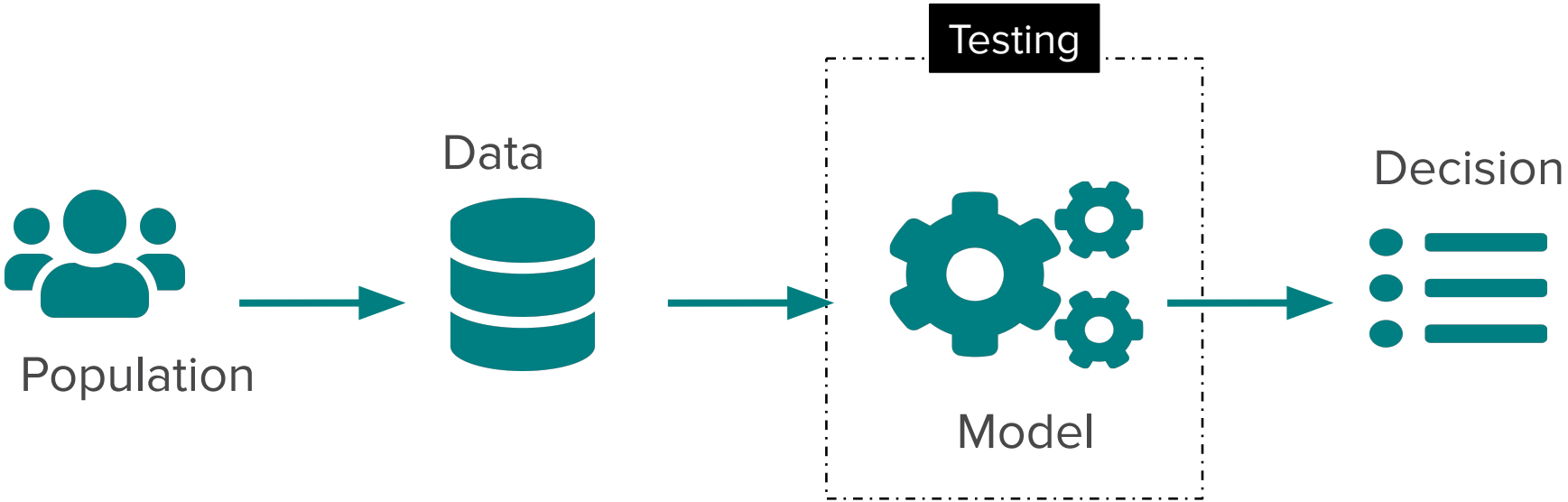


https://aif360.mybluemix.net

How do we mitigate bias in practice?

Best practice

- Check datasets imbalances
- Ensure model treats all groups fairly
- Understand model behavior
- Monitor prediction

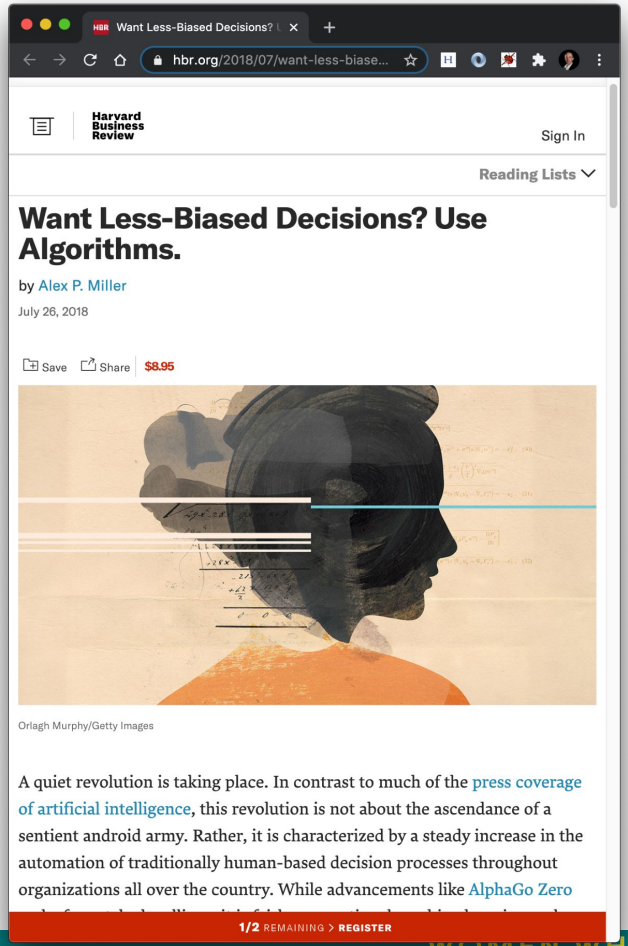


Mitigating algorithmic bias

- There is no unifying framework to tackle algorithmic bias testing and mitigation
- In most use cases, mitigation is performed after a system is built and decisions have been made based on this system

Common setting

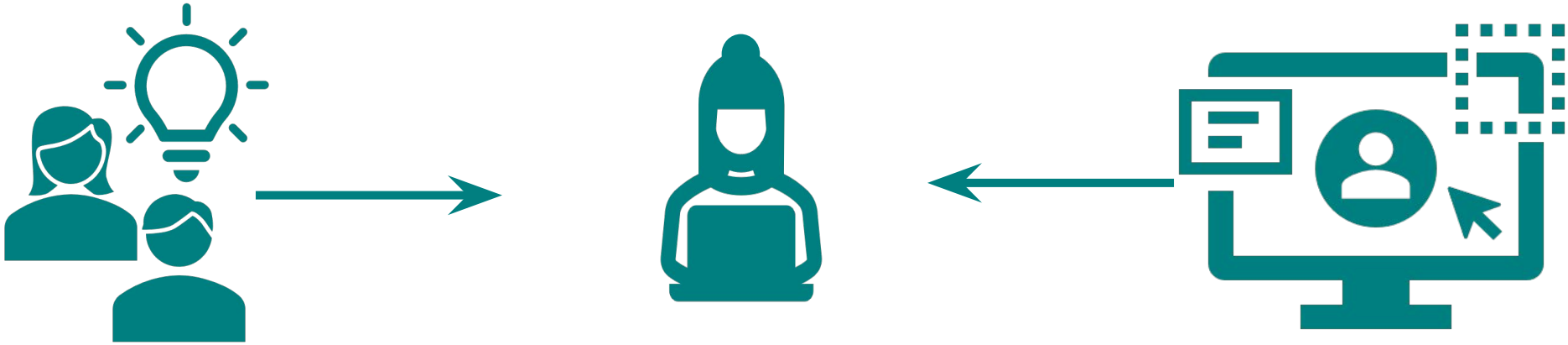
- Automate tedious or repetitive task
- Acquired or co-designed AI system
- Increasing awareness of the potential for bias



Use case: *Company acquires succession planning tool*

- **Model**: regressor trained on historical data to predict when a candidate has a positive recommendation score
- **Tool**: software that generates a promotion score for each candidate based on HR-related metrics (performance, education, tenure),
- **Practitioner task**: evaluate the tool and approve the use by HR team

Practical challenges



Business

Practitioner

Vendor

- Regulatory constraints
- Deployment/maintenance costs

- Limited agency
- Mitigation after system is built

Need for culture shift

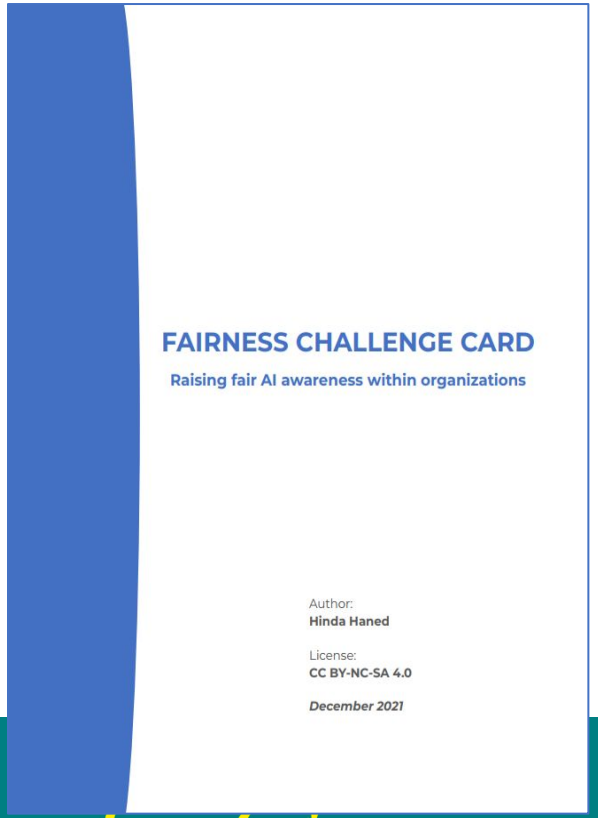
How can we make this system fair in deployment?

How do we establish SOPs to ensure fairness throughout the development process?

Fundamental questions

- Why do you need AI for this task?
- Is the system transparent?
- When and how does the system fail?
- What are the potential harms that could occur?
- What is a (un)fair outcome?
- Can we ensure fair outcomes?
- Who is responsible for ensuring fairness?

Organizational strategy for fair AI:
Raise Awareness 1/3



Questions about Privacy

The system is opaque and it's unclear to the users/customers how their data is being used and to what end

- ☐ Are the users aware of what personal data is used in the system?

No risk assessment of potentially biased outcomes.

- ☐ Are there any potentially harmful outcomes of the system identified?

GDPR test

- ☐ Can associates/users/customers opt out from using the system?
- ☐ Are the logic of the
  subjects?
- ☐ Is the impact of the
- ☐ Has there been an a
  for the intended task

Questions about Governance/Agency

A dependency on external partners who do not transfer their knowledge of the system after project completion, continuity and in-depth knowledge of the system are no longer safeguarded within the organization

- ☐ Do you understand how the system operates (Input/Output, know where to find the documentation)?

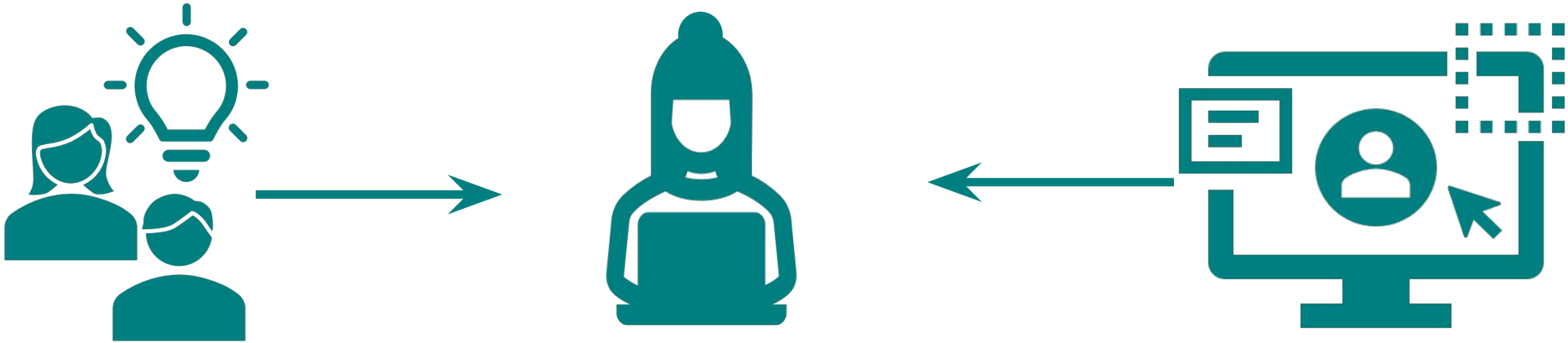
Risk of perpetuating biased/harmful outcomes with no possibility of recourse either from users or developers.

- ☐ Do you have ways to challenge the system's outputs?

Lack of accountability around model ownership and governance.

- ☐ Who is the ultimate owner (or owners) of the system?

Organizational strategy for fair AI:
Invest in talent 2/3

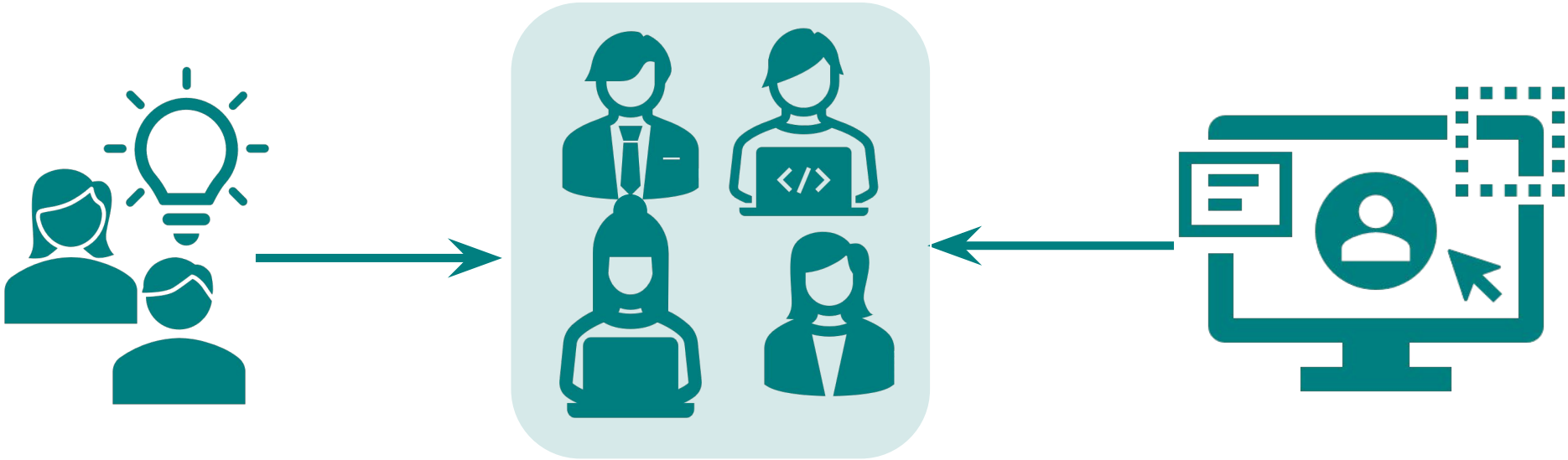


Business

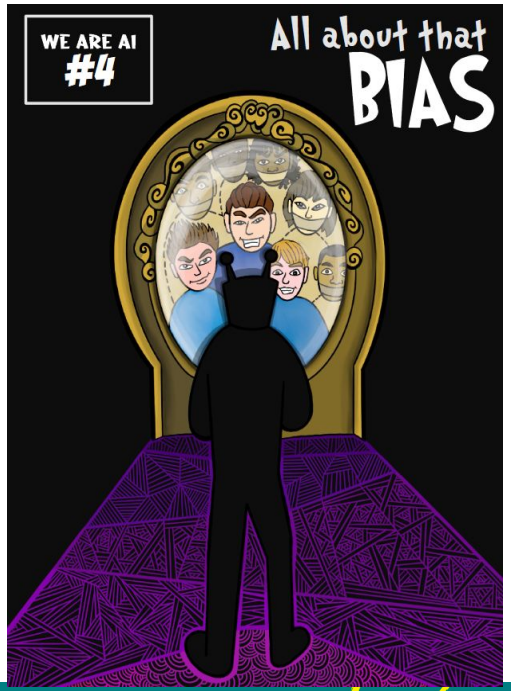
Practitioner

Vendor

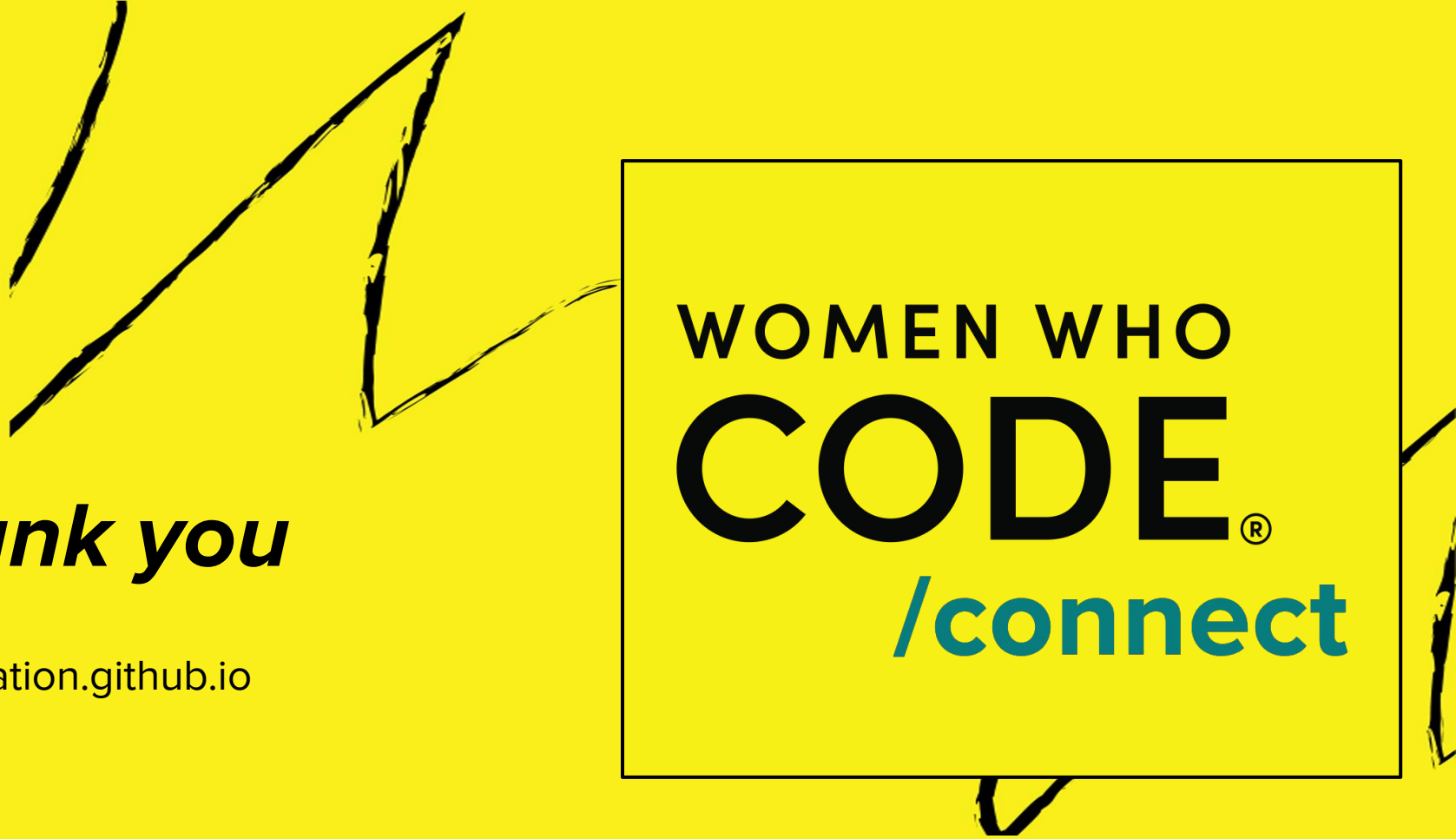
Organizational strategy for fair AI:
Invest in talent 2/3



Organizational strategy for fair AI:
Stakeholder Education 3/3



Julia Stoyanovich and Falaah Arif Khan. “All about that Bias”. We are AI Comics, Vol 4 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol4_en.pdf



Thank you

hindantation.github.io