Defining and Mitigating Algorithmic Bias

A practitioner's perspective

Hinda Haned, Ph.D.
University of Amsterdam | Owls & Arrows
August 16th, 2023

summer of AI

Background

- Entrepreneur | Data Science ~ 2022
- Endowed professor data science ~ 2018
- Lead data scientist retail/pharma ~ 2015-2020
- Forensic statistician ~ 2010 - 2015
- PhD applied statistics ~ 2010

Practitioner’s perspective

Practitioner: Anyone who needs to take a decision based on an automated (AI) system OR/AND must answer questions about possible harms caused by the system: data analysts/scientists, business analysts, business leaders, policy/compliance officers

Responsible AI

A set of best practices, guidelines, and tools that ensure any AI-driven is trustworthy, safe, and respectful of human rights and dignity

Algorithmic bias

Systematic errors of an AI system can cause significant harm to individuals and communities

Algorithmic bias

Algorithmic bias

- Biased data
- Unclear tasks
- Flawed model design
- Stereotypes
- Opaque systems

Responsible AI: why bother?

8

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?

Harmful outcomes

FROM POLITICOPRO

BY MELISSA HEIKKILÄ

MARCH 29, 2022 | 6:14 PM

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

Compliance


EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

The EU AI ACT

Ethics Guidelines

Building trust

Responsible AI in practice

How do we avoid algorithmic bias ?

slido.com
#2031

Responsible AI pillars

Fairness

“AI systems” should treat people fairly, that is, without discrimination on the grounds of protected sensitive characteristics such as age, gender, disability, ethnic or racial origin, religion or belief, or sexual orientation

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)

Harm of allocation

FROM POLITICOPRO

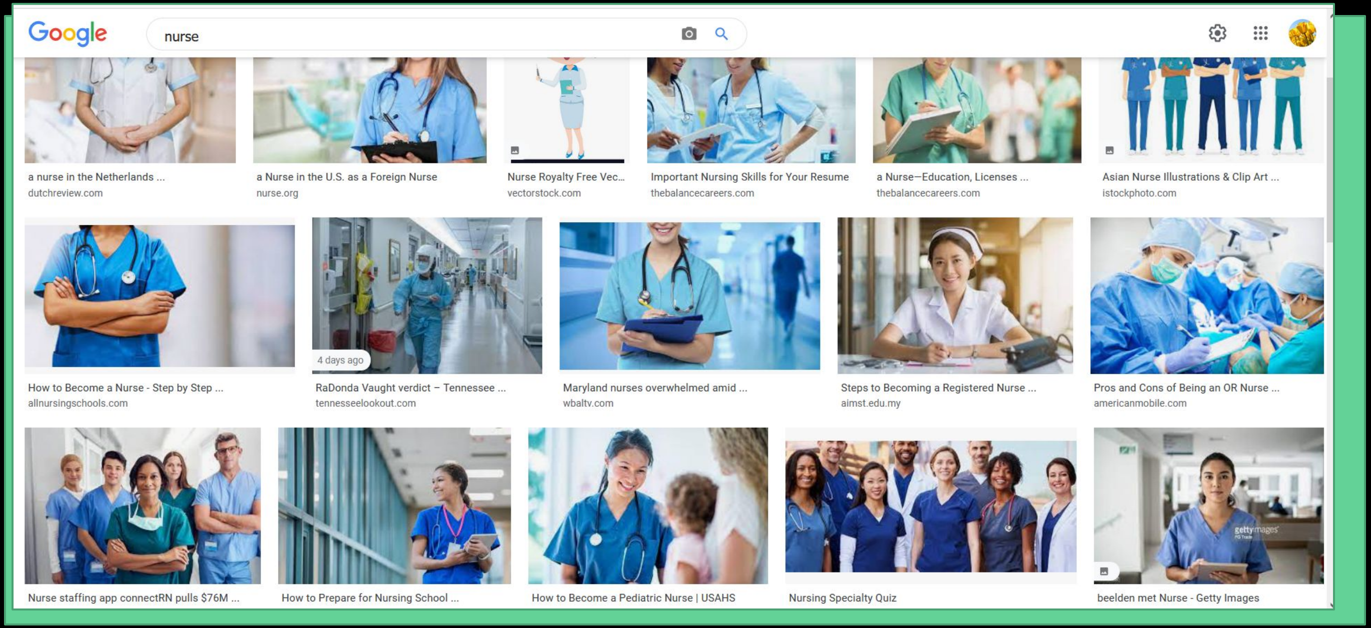
BY MELISSA HEIKKILÄ

MARCH 29, 2022 | 6:14 PM

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

Harm of representation



Queried APR22

Fairness testing

Goal: different groups experience comparable outcomes; outcome is statistically independent of sensitive attribute

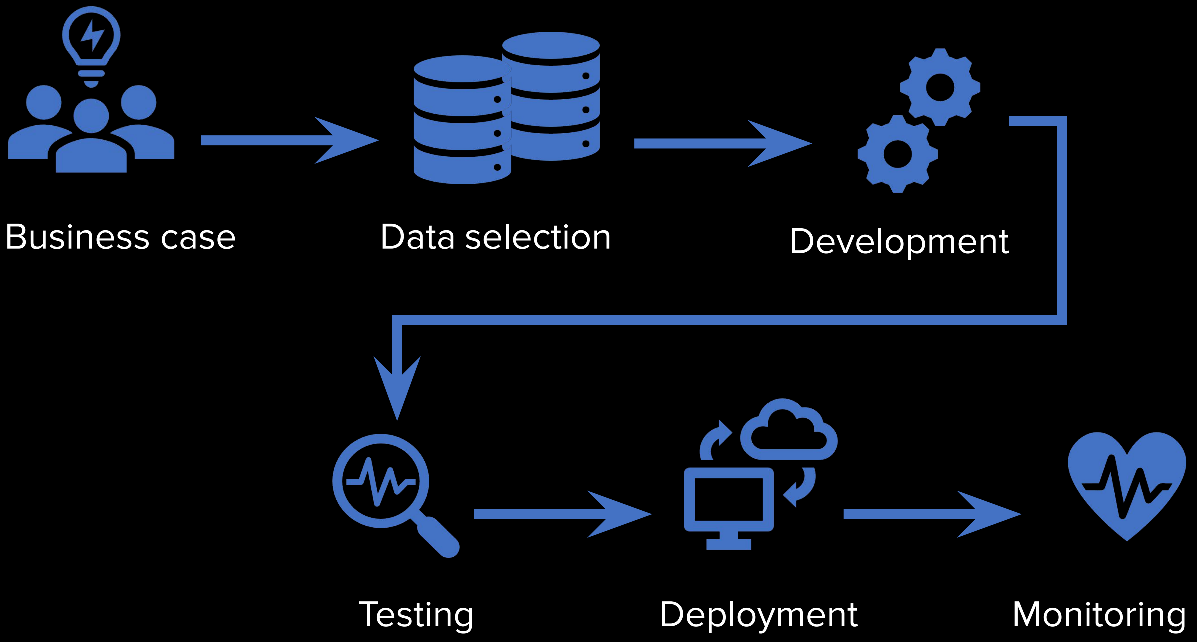
Prerequisite: sensitive attribute or group membership (e.g., age, gender, race)

Definition of fair: $E[(d(v) \mid a] = E[d(v)]$

Key insight: group-blindness does not ensure equitable group outcomes (Dwork et al., 2012)

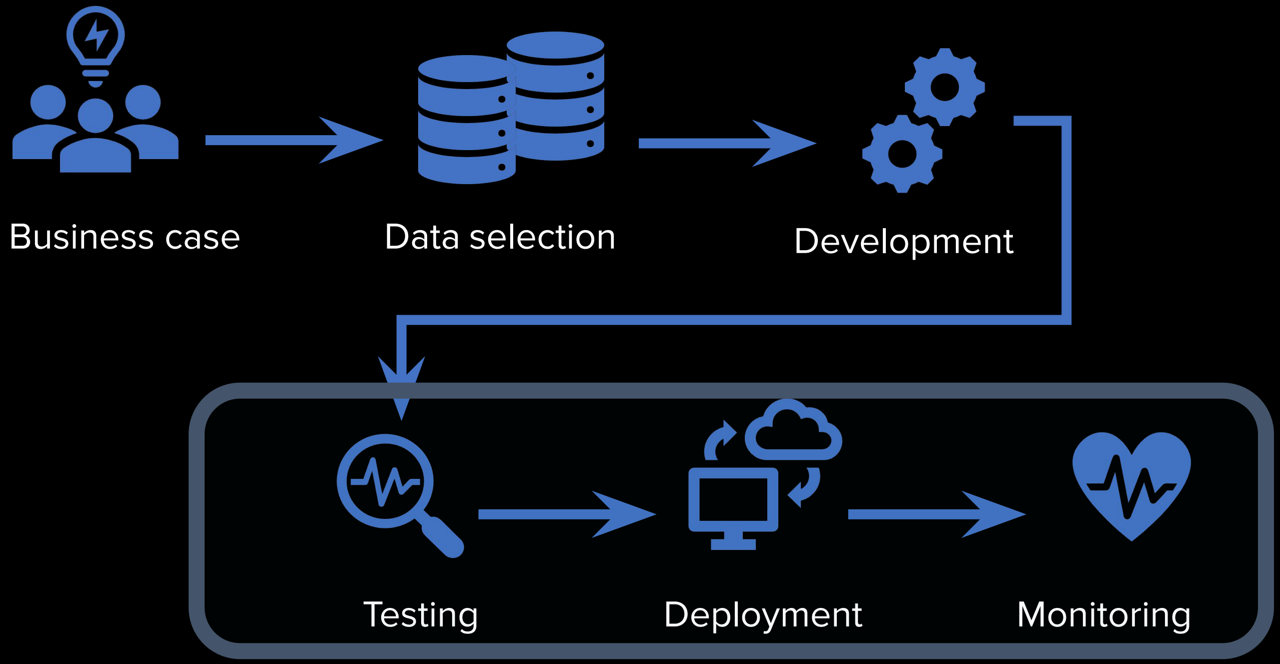
C Dwork, M Hardt, T Pitassi, O Reingold, R Zemel. Fairness through awareness. 2012

Fairness testing



Fairness testing

- Check datasets imbalances
- Ensure model treats all groups fairly



Practical challenges

In practice, there are many limitations to testing and correcting for fairness:

- Fairness testing requires unavailable/inaccessible sensitive features
- Potential fairness intervention impact cannot be monitored
- Fairness objectives not compatible with business requirements

Evaluating and mitigating algorithmic bias requires navigating uncertainty

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

Mitigation algorithms

- Mitigation **:** the action of reducing the severity, seriousness, or painfulness of something
- Mitigation algorithms: algorithms to remove or reduce bias in data and model outputs

Mitigation algorithms

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Python API Docs ↗ | Get Python Code ↗ | Get R Code ↗

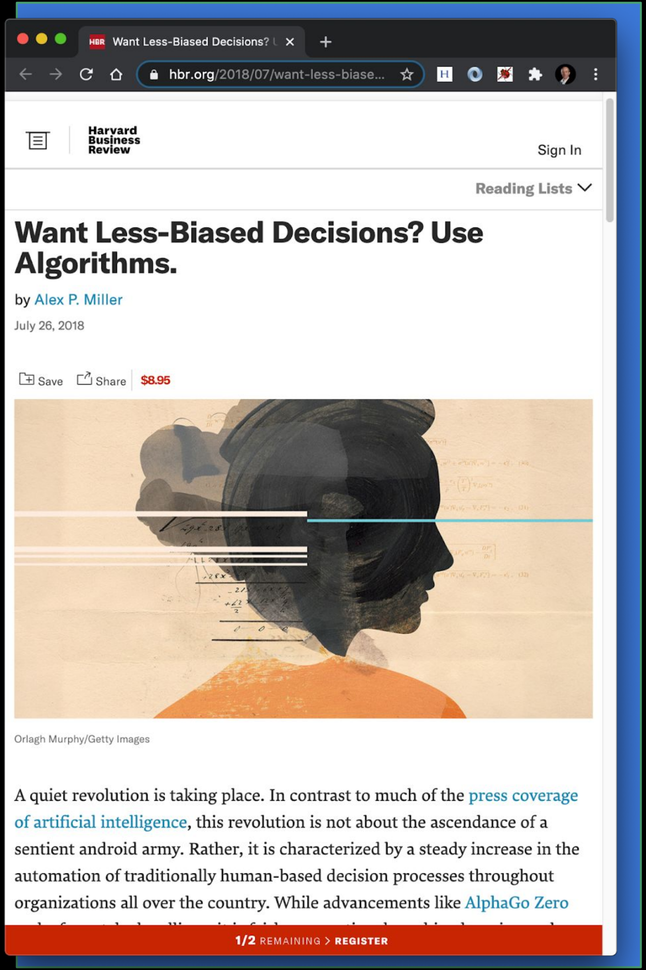
https://aif360.mybluemix.net/

- ❏ Known sensitive attributes
- ❏ Defined fairness objectives
- ❏ Fairness intervention
- ❏ Monitor fairness intervention

Use case

Common setting

- Automate tedious or repetitive task
- AI System acquired or co-designed
- Challenged by end-user adoption and acceptance


Want Less-Biased Decisions?
hbr.org/2018/07/want-less-biase...
Harvard Business Review
Sign In
Reading Lists

Want Less-Biased Decisions? Use Algorithms.

by Alex P. Miller

July 26, 2018

Save Share $8.95

Orlagh Murphy/Getty Images

A quiet revolution is taking place. In contrast to much of the press coverage of artificial intelligence, this revolution is not about the ascendance of a sentient android army. Rather, it is characterized by a steady increase in the automation of traditionally human-based decision processes throughout organizations all over the country. While advancements like AlphaGo Zero

1/2 REMAINING > REGISTER

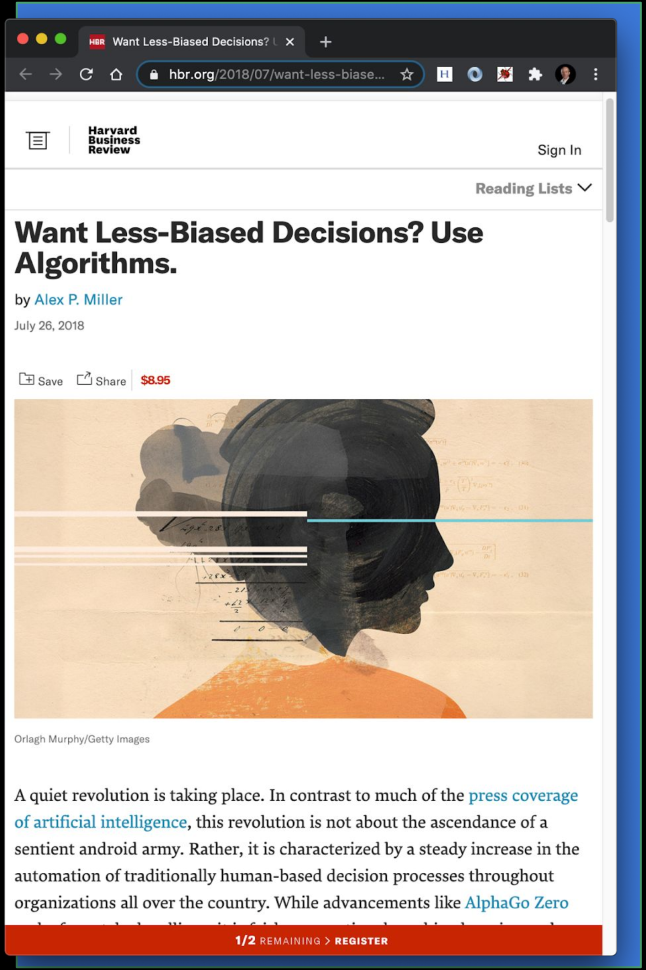
Use case: Company acquires succession planning tool

- **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

Common setting

- Automate tedious or repetitive task
- AI System acquired or co-designed
- Challenged by end-user adoption and acceptance

How can we make this system fair in deployment?



Want Less-Biased Decisions? U
hbr.org/2018/07/want-less-biase...

Harvard Business Review

Sign In

Reading Lists

Want Less-Biased Decisions? Use Algorithms.

by Alex P. Miller

July 26, 2018

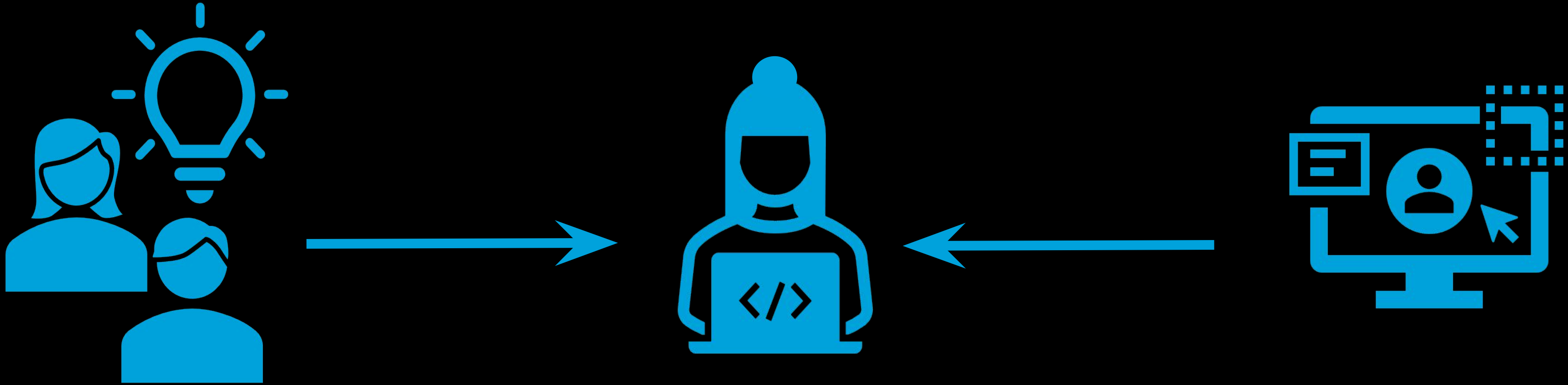
Save Share $8.95

Orlagh Murphy/Getty Images

A quiet revolution is taking place. In contrast to much of the press coverage of artificial intelligence, this revolution is not about the ascendance of a sentient android army. Rather, it is characterized by a steady increase in the automation of traditionally human-based decision processes throughout organizations all over the country. While advancements like AlphaGo Zero

1/2 REMAINING > REGISTER

Limited agency



Business

Practitioner

Vendor

- Regulatory constraints
- Deployment/maintenance costs

- Limited agency
- Mitigation after system is built

Need for a culture shift

How can we make this system fair in deployment?

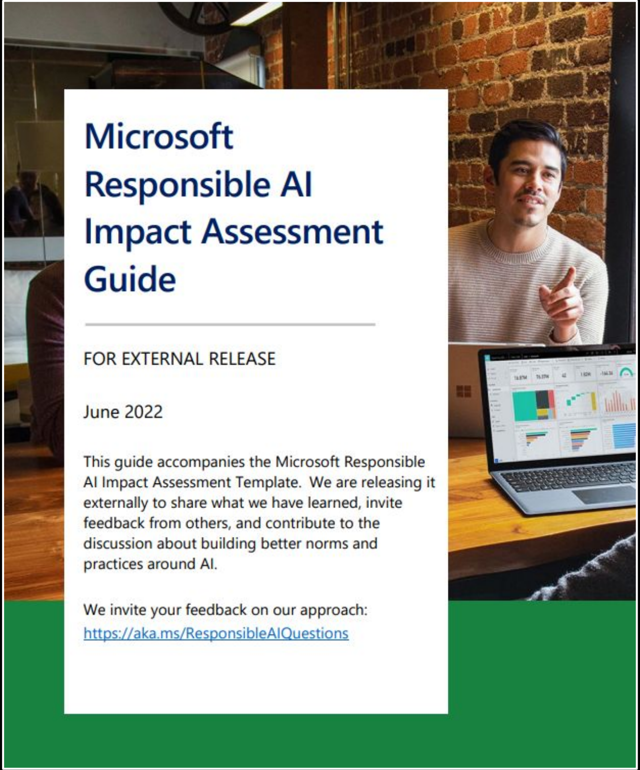
How do we establish SOPs to ensure reliability, transparency & accountability throughout the development process?

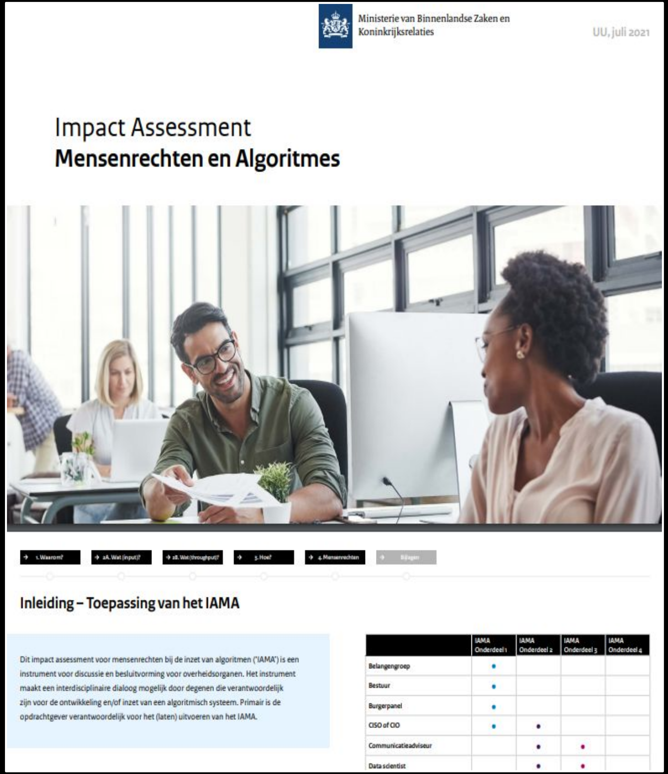
Ask fundamental questions

- 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?

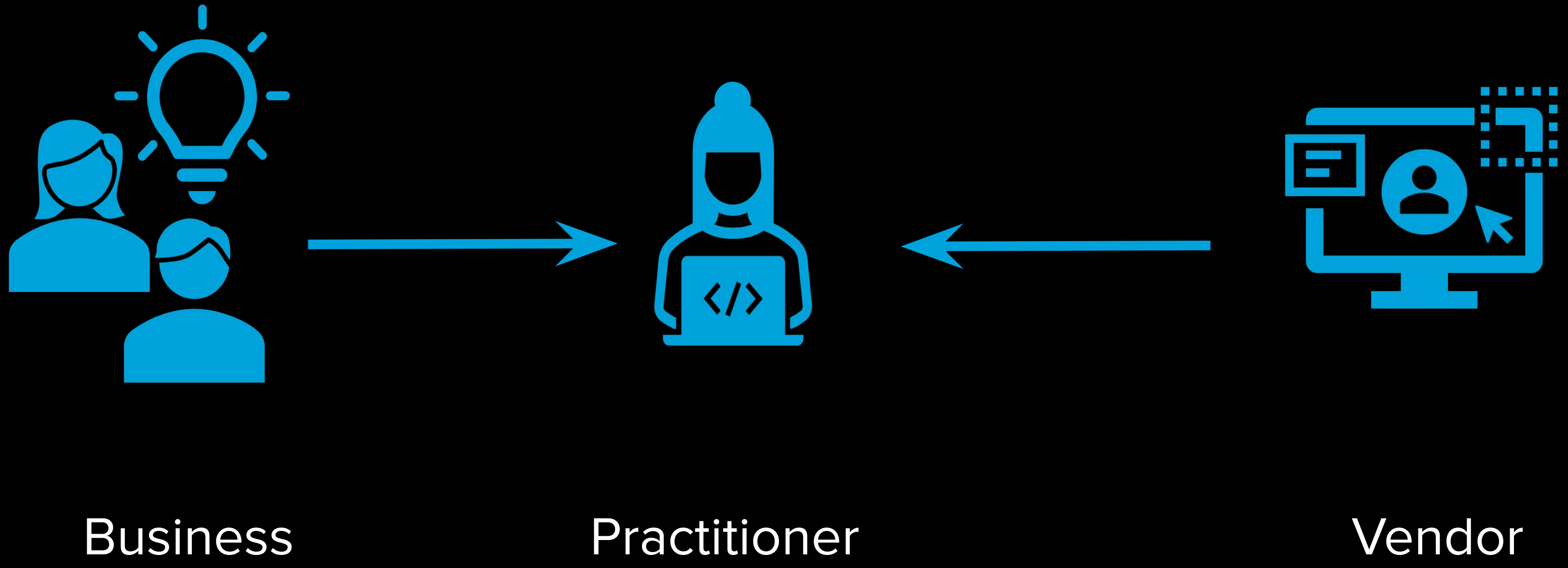
34

Perform impact assessments



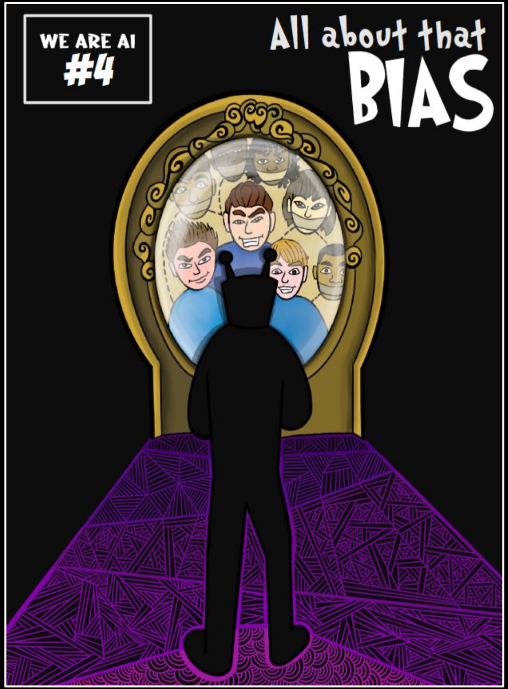


Invest in talent



Invest in talent

Educate stakeholders



Julia Stoyanovich and Falaah Arif Khan. “All about that Bias”. We are AI Comics, Vol 4 (2021)

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

38

Adopt best practices

Reproducibility

User centricity

Peer review

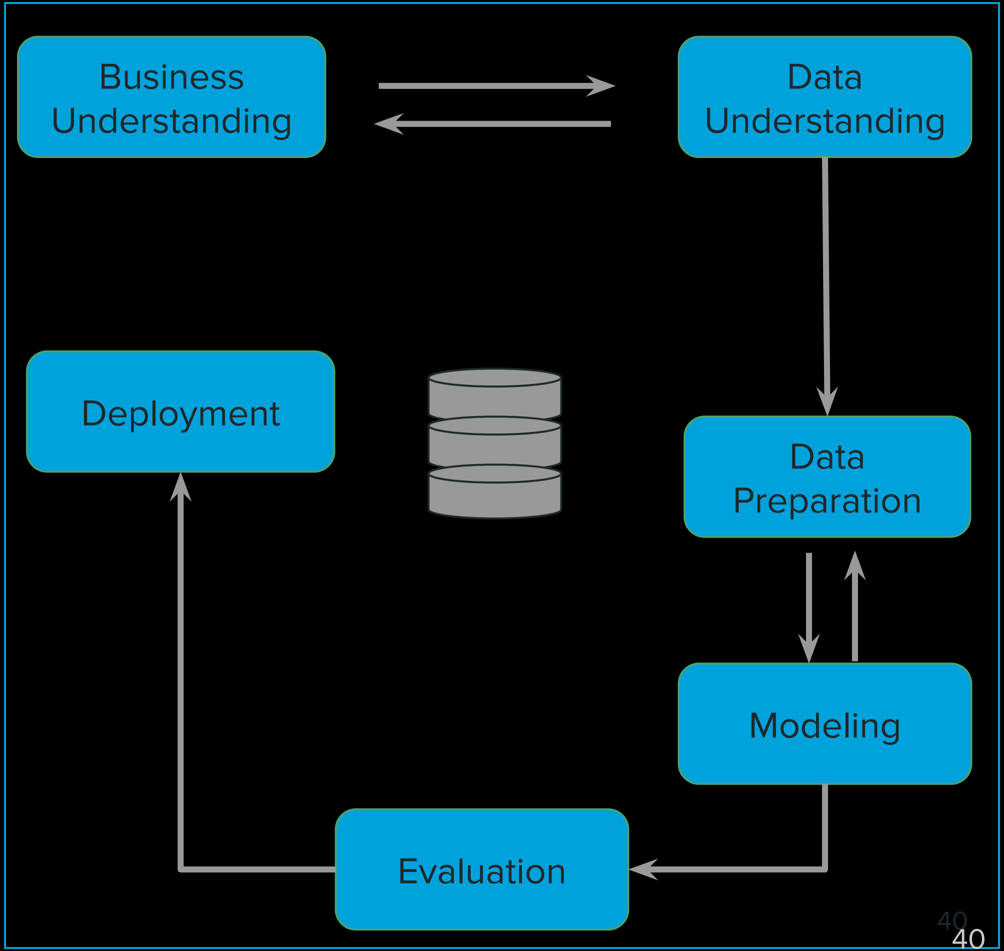
Transparency

39

Adopt best practices

Example: CRoss Industry Standard Process for Data Mining (CRISP-DM)

Adapted from: https://www.datascience-pm.com/crisp-dm-2/



Consider AI technology an ecosystem

What can YOU do?

Raise Awareness

Provide Use-cases

Monitor AI Applications

Conduct (internal) Audits

Drive Stakeholder Centered R&D

Influence Policies & Regulations

Thank you!

h.haned@uva.nl
hindantation.github.io

Owls & Arrows

Fairness challenge card

FAIRNESS CHALLENGE CARD

Raising fair AI awareness within organizations

Author:
Hinda Haned

December 2021

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?