EPFL

Decision making game: debriefing

CS-433

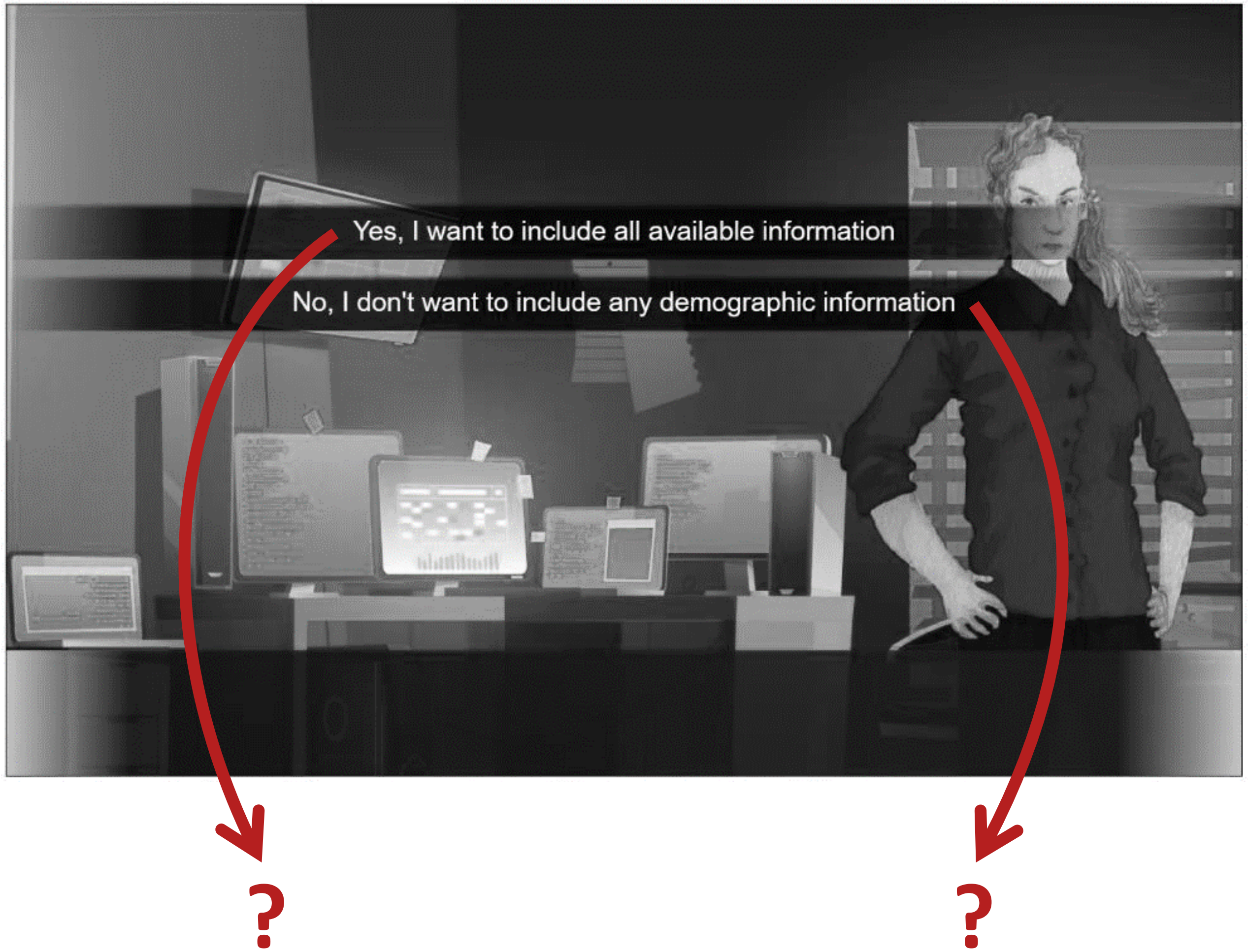
21 November 2023

Cécile Hardebolle
EPFL Center for Digital Education
cecile.hardebolle@epfl.ch

Let’s refresh our memory

In the prisoner case:

- What decision did you take: include the demographic information or not ?
- What was the consequence?



Yes, I want to include all available information

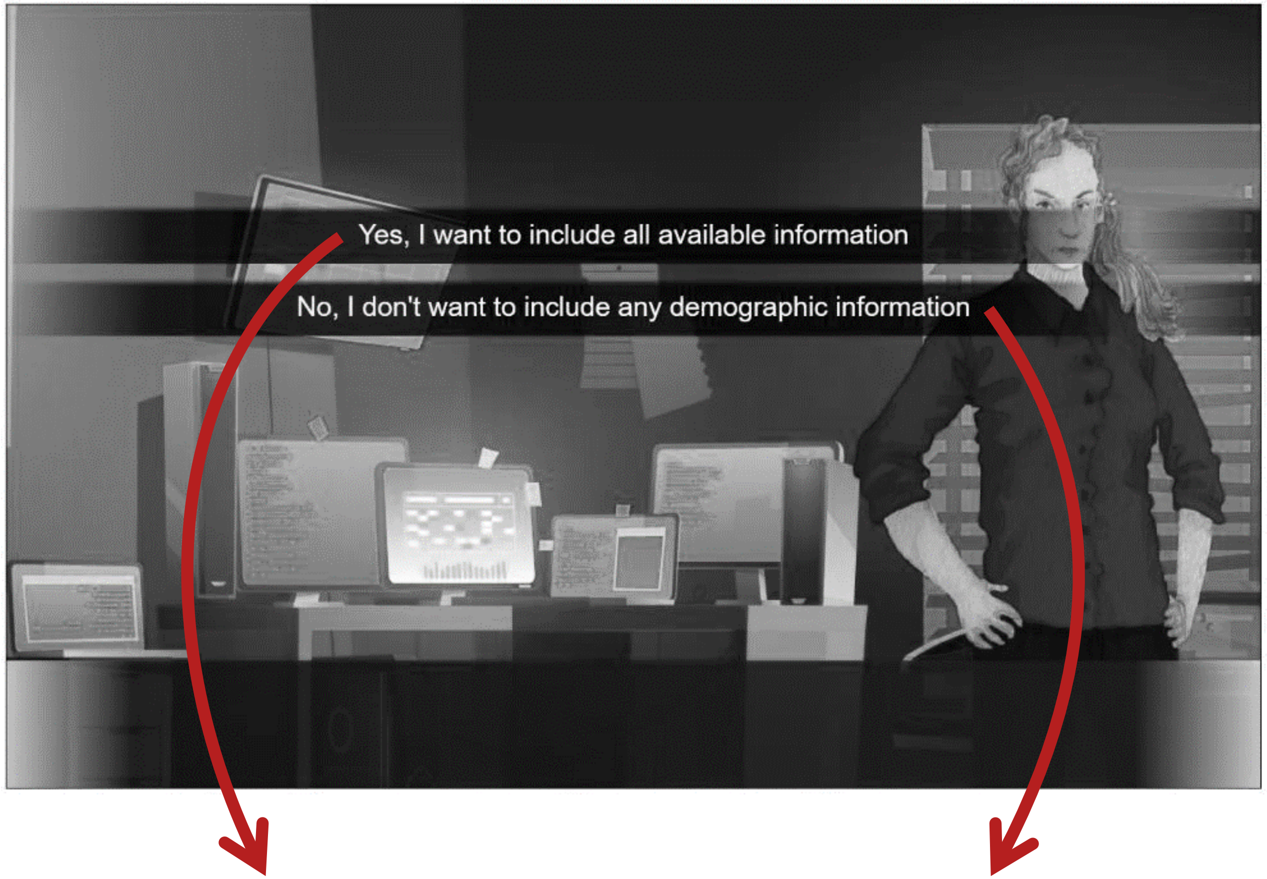
No, I don't want to include any demographic information

Fairness vs. Accuracy

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. ACM Computing Surveys, 55(3), 51:1-51:44. https://doi.org/10.1145/3494672

In the prisoner case:

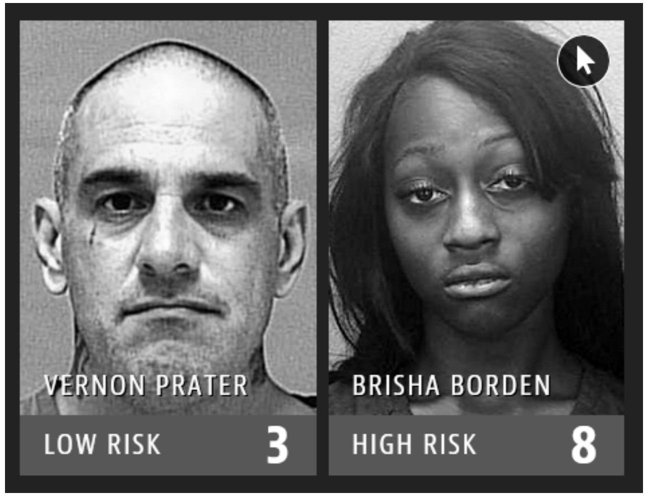
- What decision did you take: include the demographic information or not ?
- What was the consequence?

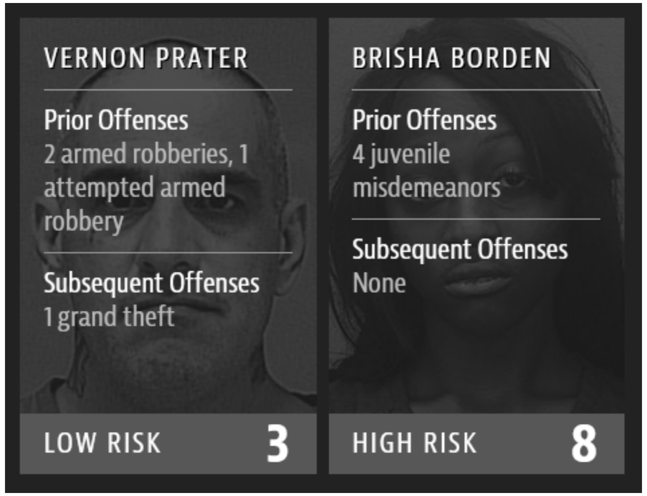

Yes, I want to include all available information
No, I don't want to include any demographic information

Racist model
▶ Friend not released

Low accuracy
▶ Neighbor robbed

Based on real cases

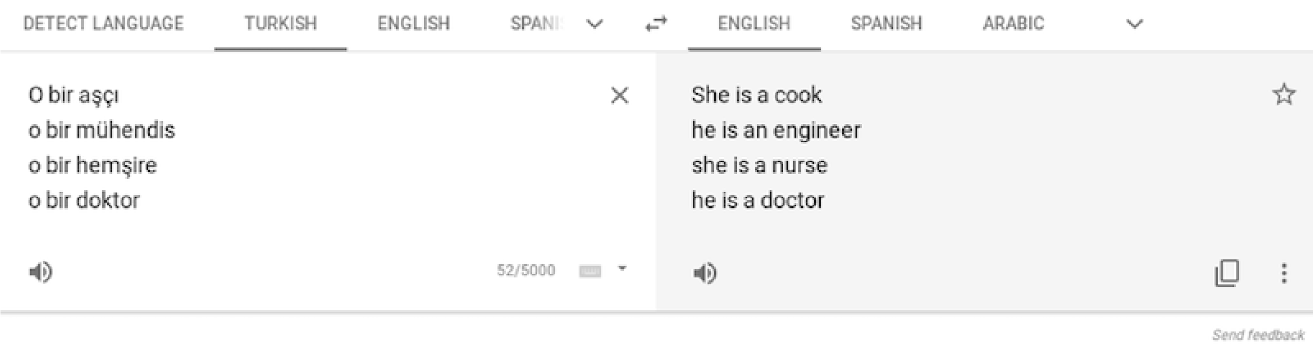




VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias : There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

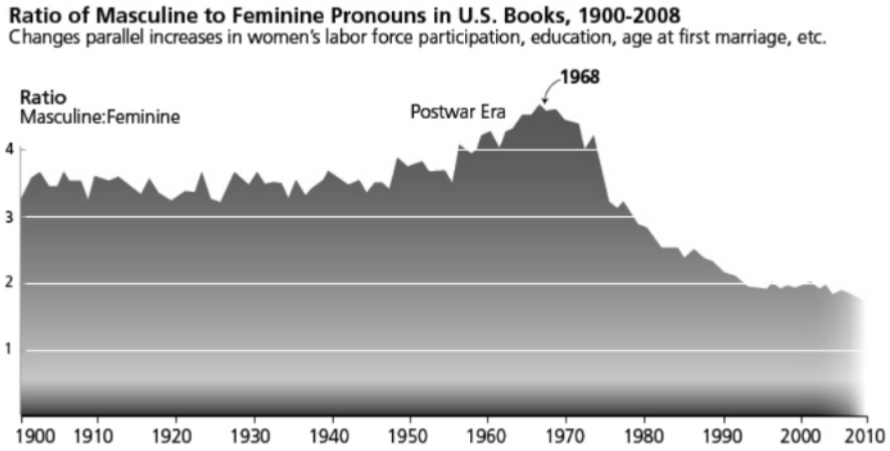
Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580



O bir aşçı
o bir mühendis
o bir hemşire
o bir doktor

She is a cook
he is an engineer
she is a nurse
he is a doctor

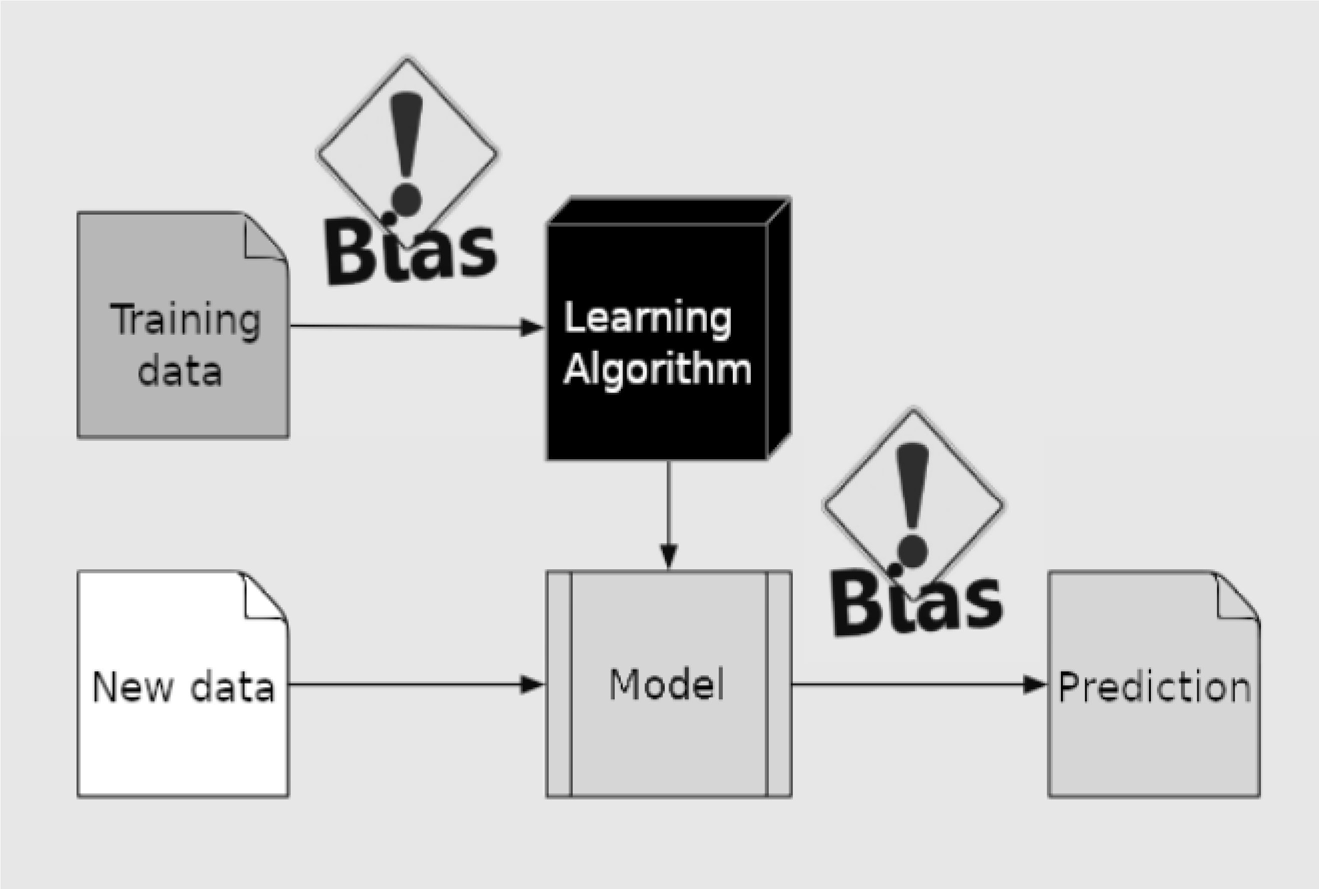
Machine Translation | Gendered Innovations
http://genderedinnovations.stanford.edu/case-studies/nlp.html



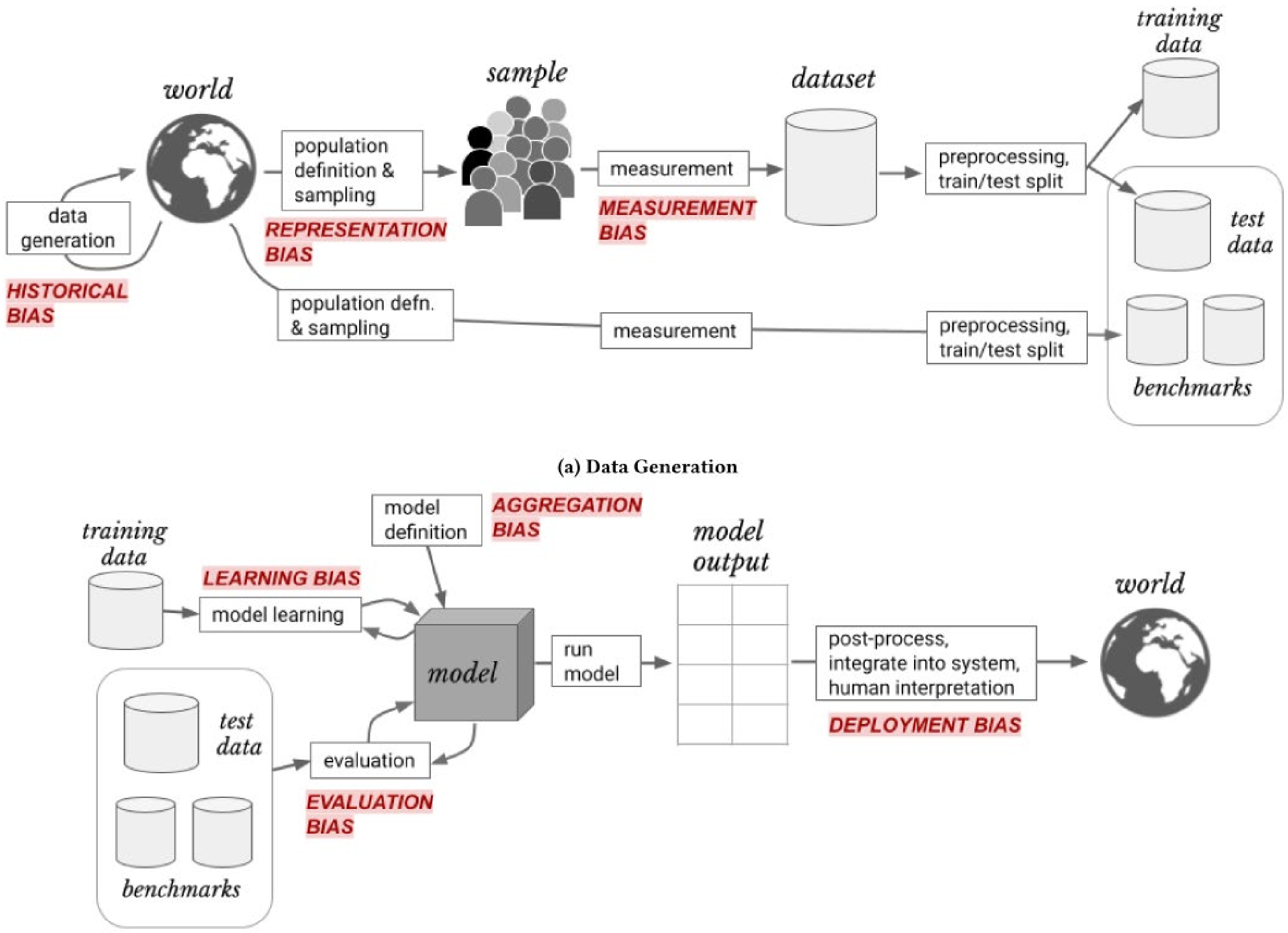
The ratio of masculine pronouns ("he," "him," "his," "himself") to feminine pronouns ("she," "her," "hers," "herself") peaked at over 4:1 in 1968. By 2000 the ratio dropped dramatically to 2:1 (Twenge et al., 2012).

Data from American English corpus of the Google Books database (~1.2 million books).
Reproduced from Twenge et al., 2012.

Bias and fairness in Machine Learning



Bias and fairness in Machine Learning



(a) Data Generation

(b) Model Building and Implementation

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

Most Machine Learning systems will affect people

“Crowd” by Amy West on Flickr, CC BY 2.0
https://www.flickr.com/photos/amy_elizabeth_west/3876549126/

Risk: harm at **scale**

FROM POLITICO PRO

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.

As the world turns to AI to automate their systems, the Dutch scandal shows how devastating they can be

The Dutch Tax Authority Was Felled by AI—What Comes N...

NEWS | ARTIFICIAL INTELLIGENCE

The Dutch Tax Authority Was Felled by AI—What Comes Next? › European regulation hopes to rein in ill-behaving algorithms

BY RAHUL RAO | 09 MAY 2022 | 4 MIN READ

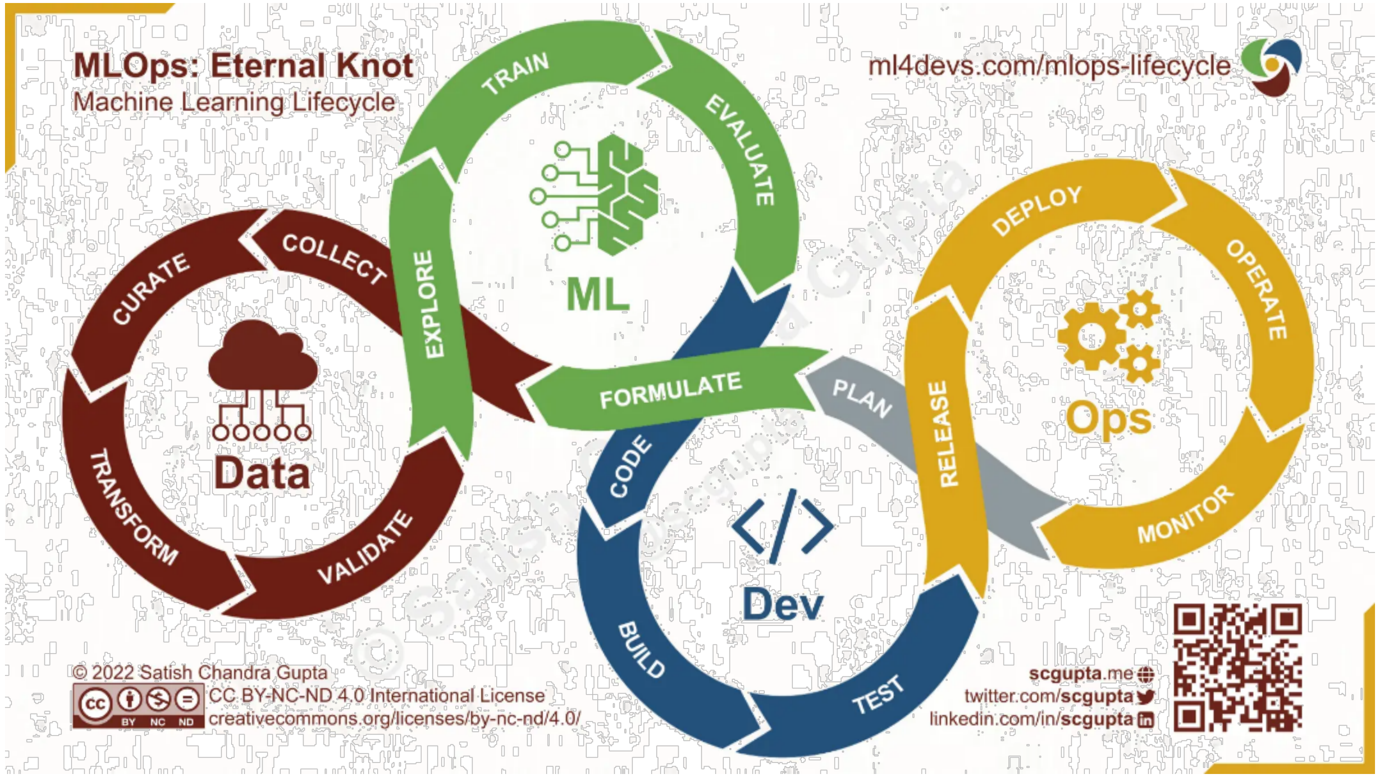
October 25, 2021

Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms

Recently added

Pakistan: Ban on film Joyland showcasing transgender character

Your decisions can make a difference!



"It is important to acknowledge that **not all problems should be blamed on the data**. The ML pipeline involves a series of choices and practices, from model definition to user interfaces used upon deployment. **Each stage involves decisions that can lead to undesirable effects.**"

(Suresh & Guttag, 2021)

Responsible design decisions

→ Evaluate **who could be affected** and **how**

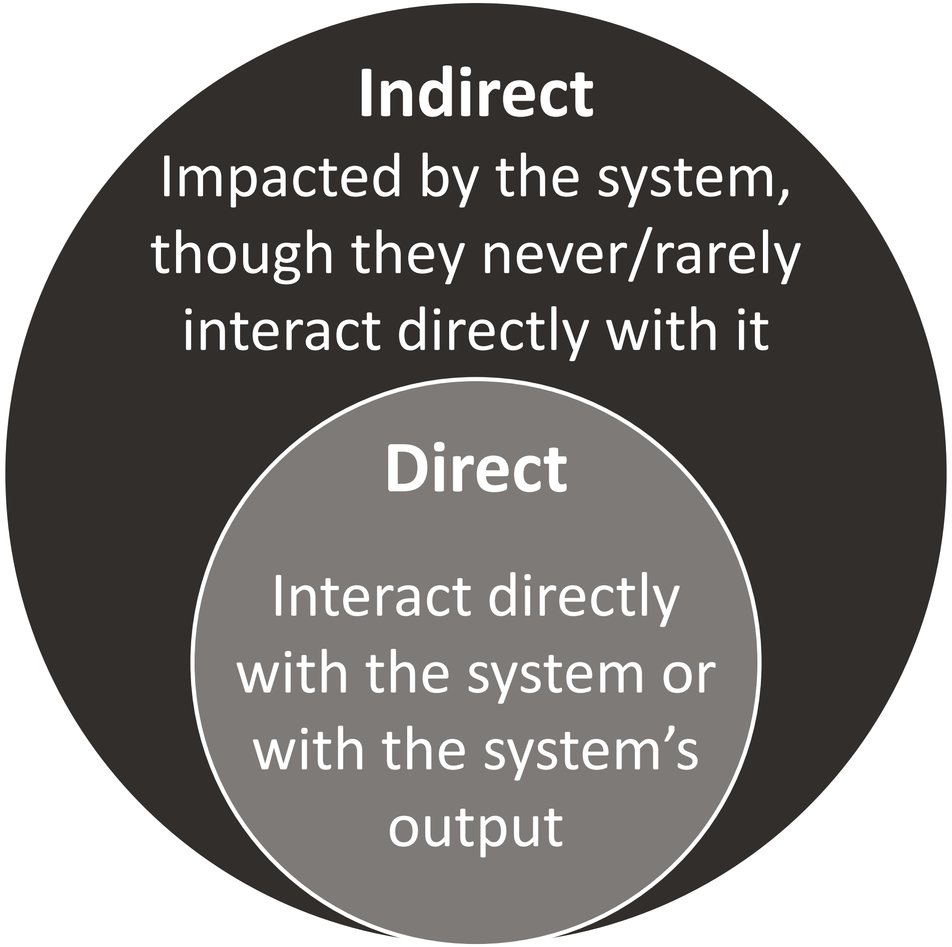
2 strategies to practice in your project!

↳ **Required “Ethical risks” section in your report**

Responsible design decisions

→ Evaluate **who could be affected** and **how**

Stakeholder Analysis



Indirect

Impacted by the system, though they never/rarely interact directly with it

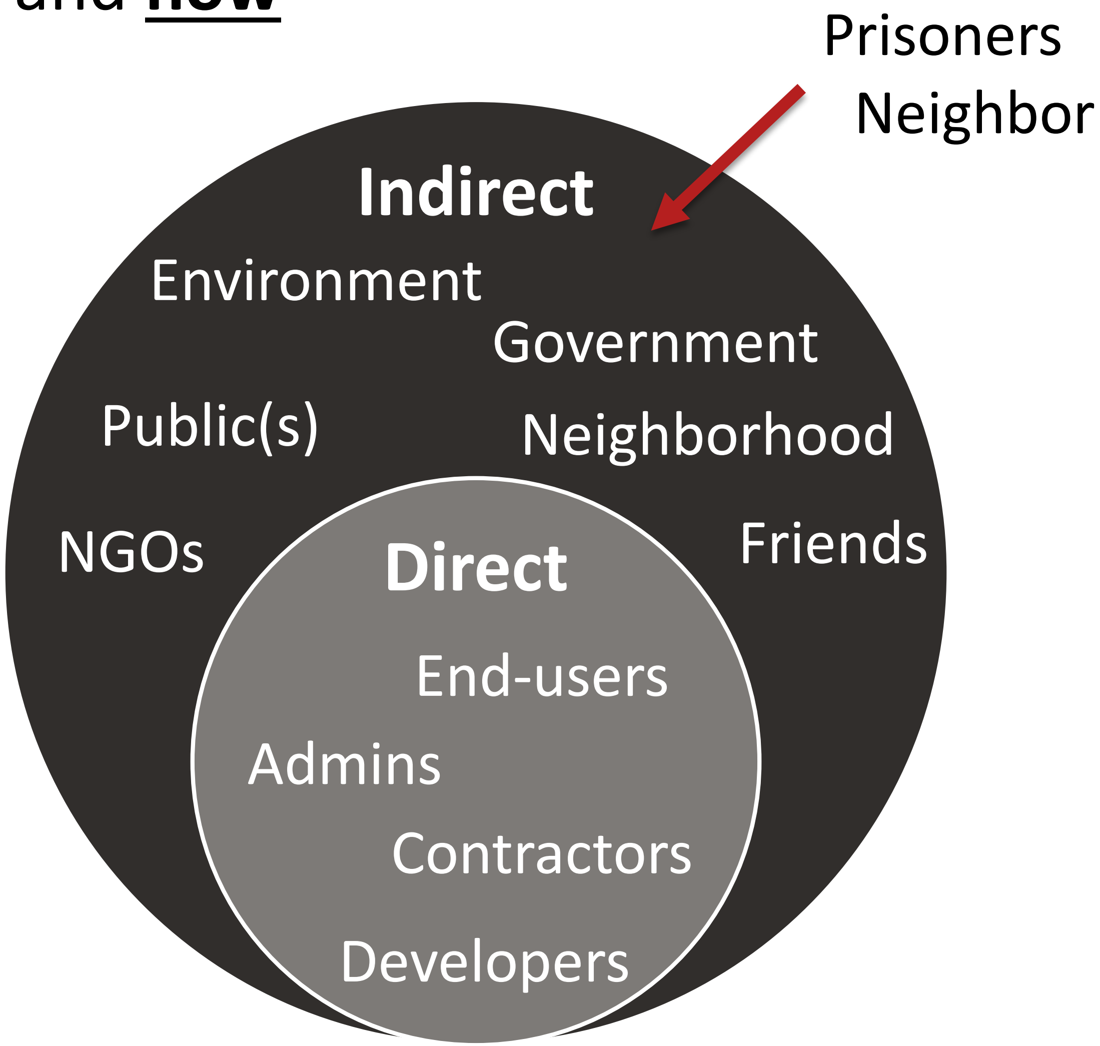
Direct

Interact directly with the system or with the system’s output

Responsible design decisions

→ Evaluate **who could be affected** and **how**

Stakeholder Analysis



Responsible design decisions

→ Evaluate **who could be affected** and **how**

1. Speculative scenarios
2. Research + measure

Risk Analysis

Negative impact

+ Likelihood of occurrence

+ Severity

Case: ML model that predicts user emotions based on smartphone touch data

Imagine it is **deployed in a social media platform**:

- Included in the user interface to display predicted emotions
- Accessible to third parties (e.g. for ads)
- Used for internal functions (e.g. content recommendation or moderation)

What could go wrong? Generate as many ideas as possible:

□ Can the solution be used in **harmful ways**?

□ What kind of impacts can **errors** from the solution have?

□ Could the solution disclose / be used to **disclose private information**?

□ Could the solution contribute to **discrimination** against people or groups?

□ What **choices** are users able to make in their use of the solution and how?

Context

□ In which context is the solution evaluated?

Beneficence

□ What are the expected benefits of the solution in this context?

Non-maleficence

Risks	Mitigation
□ Can the solution be used in harmful ways, in particular with regards to vulnerable populations? □ What kind of impacts can errors from the solution have? □ What type of protections does the solution have against attacks or misuse?	

Solution

□ What are the characteristics of the solution under evaluation?

Privacy

Risks	Mitigation
□ What data does the solution collect? □ Is it collecting personal or sensitive data? □ Who has access to the data? □ How is the data protected? □ Could the solution disclose / be used to disclose private information?	

Fairness

Risks	Mitigation
□ How accessible is the solution? □ What kinds of biases may affect the results? □ Can the outcomes of the solution be different for different users or groups? □ Could the solution contribute to discrimination against people or groups?	

Sustainability

Risks	Mitigation
□ What is the carbon footprint of the solution? □ What types of resources does it consume (e.g. water) and produce (e.g. waste)? □ What type of human labor is involved?	

Empowerment

Risks	Mitigation
□ Can users understand how the solution works and what its limits are? □ Are users able to make choices (e.g. consent, settings) in their use of the solution and how? □ How does the solution affect user autonomy and agency?	

Have you identified a range of different types of risks?

Context

□ In which context is the solution evaluated?

Beneficence

□ What are the expected benefits of the solution in this context?

Non-maleficence

Mitigation

Generate mis-communication and conflict

Use for scams

Use for bullying

…e used in harmful …ith regards to …ations? …npacts can errors from the … □ …tections does the s… …nst attacks or misuse?

Privacy

Mitigation

People lose control over private emotions

Mental health data

Possible to identify political opinions

□ … solution collect? …onal or sensitive …ss to the data? □ H… …otected? □ C… disclose / be used to dis… …rmation?

Fairness

Risks	Mitigation
Bias wrt cultural differences Bias wrt people with handicap (physical / mental)	

□ … is the solution? □ …ases may affect the r… □ Can the outcomes of the solution be … users or groups? … contribute to … …nst people or groups?

Sustainability

Risks	Mitigation
CO2 & water for training + inference Emotion labeling for training	

What is the carbon footprint of the … …esources does it consume (e.g. water) and produce (e.g. … …uman labor is involved?

Empowerment

Mitigation

People lose control over private emotions

People stop relying on their own interpretation

Attention capture, manipulation

…stand how the solution … limits are? □ A… …e choices (e.g. con… …r use of the solu… □ …lution affect user a… …ncy?

Solution

□ What are the characteristics of the solution under evaluation?

User guide and examples on https://go.epfl.ch/canvas

You can make a difference!

ML systems can have **harmful consequences for people**

↳ Evaluate **who** could be affected and **how:**

- Stakeholder analysis
- Ethical risk analysis

https://go.epfl.ch/canvas

↳ Document the **ethical risk analysis** you have performed

In your project report:

- Describe 1 type of risk (or justify the absence of risks)
- Explain how you evaluated it
- Describe how you took it into account (or the barriers to do it)

Ask your questions on Ed (tag me)