REVISION WEEK

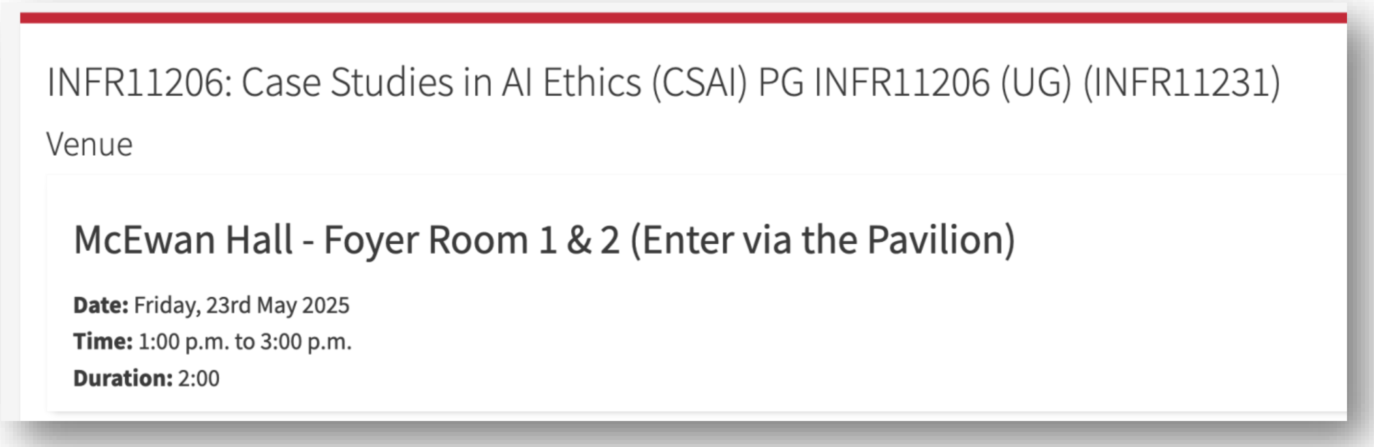
Week 1 - Week 9

Exam
to do list
-~~brainstorm~~
-~~research~~

EXAM LOGISTICS

- https://exams.is.ed.ac.uk/

(check here in case there are any changes)



INFR11206: Case Studies in AI Ethics (CSAI) PG INFR11206 (UG) (INFR11231)

Venue

McEwan Hall - Foyer Room 1 & 2 (Enter via the Pavilion)

Date: Friday, 23rd May 2025
Time: 1:00 p.m. to 3:00 p.m.
Duration: 2:00

EXAM STRUCTURE

- Three Questions
  - Question 1 is compulsory
  - ONE other question
- Closed book!
- Questions cover:
  - Data Ethics
  - Machine Ethics
  - AI Ethics

EXAM STRUCTURE

- Expect to have small case studies
- Ensure you have done all the required readings for the course
- Ensure you know about all the materials covered during the lectures/tutorials
- You are not responsible for the guest lectures

READ CAREFULLY, BE BRIEF, JUSTIFY YOUR RESPONSES!

EXAMPLE QUESTION - AI ETHICS

Company X decided to make an internal audit for one of its client projects. Happy-Go-Lucky, Inc., an imagined photo service company looking for a smile detection algorithm to automatically trigger the cameras in their installed physical photo booths [1]. Company X has designated five AI principles: "Transparency", "Justice, Fairness & Non-Discrimination", "Safety & Non-Maleficence", "Responsibility & Accountability" and "Privacy".

[1] Use case is adapted from the following paper: Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44.

Q1

In class, we discussed various AI principles that an organization may want to follow. Consider Company X’s AI principles, and answer the questions.

(a) State the definition of one of the AI principles.

(b) Name one AI principle that Happy-Go-Lucky could violate. Justify your response.

(c) Name one solution to mitigate this problem.

Q2

Use Failure Modes and Effects Analysis (FMEA), methodical and systematic risk management approach to examine Happy-Go-Lucky company.

(a) Choose one specific feature with low risk priority and apply the methodology.

(b) Choose one specific feature with high risk priority and apply the methodology.

Q3

As Company X continues its internal audit, it gets access to the dataset that Happy-Go-Lucky is using to train their smile detection algorithm. The artifact A includes the following demographic details: (58.1% female, 42% male), (77.8% aged 0-45, 22.1% aged over 46 and 14.2% lighter-skinned, 85.8% darker-skinned).

(a) What does artifact A represent?

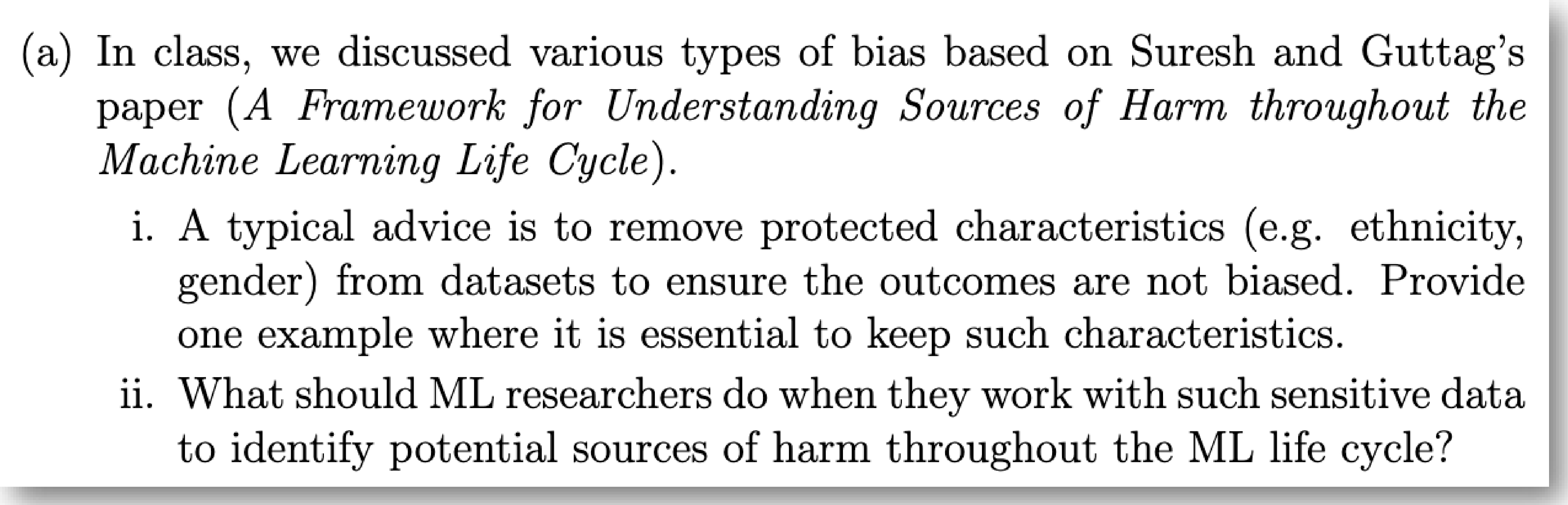
(b) By looking at A, what is the AI principle Happy-Go-Lucky violates? Justify your response.

(c) What mitigation strategy could be recommended to address the problem in 3.(b)? How would you ensure that this strategy is good?

Q4 -- MACHINE ETHICS

- Expect questions based on readings/lectures.
- Ignore all machine ethics questions from previous years

Q5 (CSAI EXAM 21-22) -- READINGS

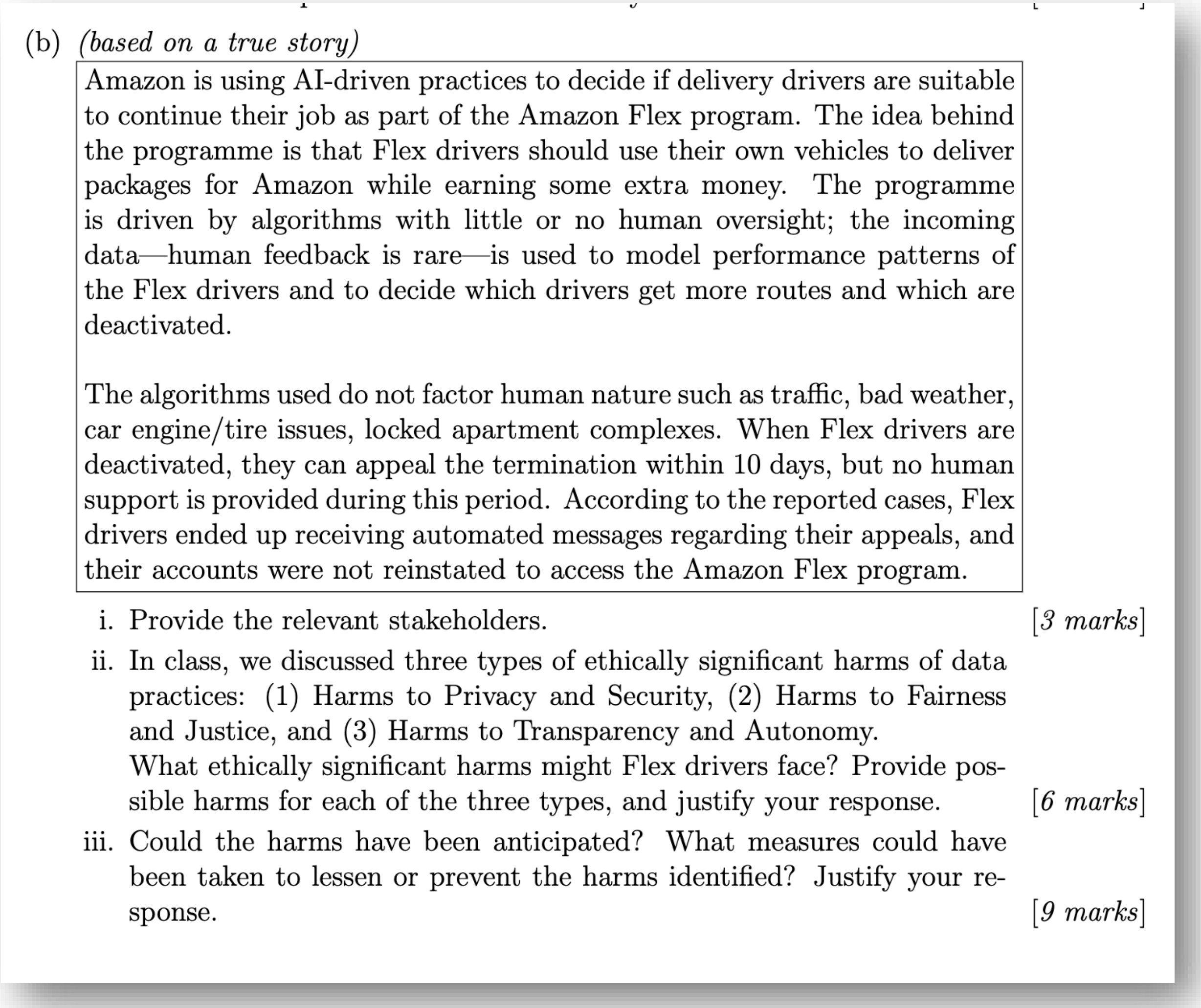


(a) In class, we discussed various types of bias based on Suresh and Guttag's paper (*A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*).

i. A typical advice is to remove protected characteristics (e.g. ethnicity, gender) from datasets to ensure the outcomes are not biased. Provide one example where it is essential to keep such characteristics.

ii. What should ML researchers do when they work with such sensitive data to identify potential sources of harm throughout the ML life cycle?

Q6 (CSAI EXAM 22-23) -- ESSAY-TYPE QUESTIONS



(b) *(based on a true story)*

Amazon is using AI-driven practices to decide if delivery drivers are suitable to continue their job as part of the Amazon Flex program. The idea behind the programme is that Flex drivers should use their own vehicles to deliver packages for Amazon while earning some extra money. The programme is driven by algorithms with little or no human oversight; the incoming data—human feedback is rare—is used to model performance patterns of the Flex drivers and to decide which drivers get more routes and which are deactivated.

The algorithms used do not factor human nature such as traffic, bad weather, car engine/tire issues, locked apartment complexes. When Flex drivers are deactivated, they can appeal the termination within 10 days, but no human support is provided during this period. According to the reported cases, Flex drivers ended up receiving automated messages regarding their appeals, and their accounts were not reinstated to access the Amazon Flex program.

i. Provide the relevant stakeholders. [*3 marks*]

ii. In class, we discussed three types of ethically significant harms of data practices: (1) Harms to Privacy and Security, (2) Harms to Fairness and Justice, and (3) Harms to Transparency and Autonomy.
What ethically significant harms might Flex drivers face? Provide possible harms for each of the three types, and justify your response. [*6 marks*]

iii. Could the harms have been anticipated? What measures could have been taken to lessen or prevent the harms identified? Justify your response. [*9 marks*]

CSAI Teaching Support Team 2025

You can be the next?

- Sydelle de Souza (TA)
- Fiona Smith (Tutor)
- Passara Chanchotisatien (Marker)
- Sunnie Li (Marker)

Edinburgh
University
**Students'
Association**

Nominations close on 28 March, 10am

https://www.eusa.ed.ac.uk/whatson/awards/teachingawards

REVISION

What makes a harm or benefit ethically significant?

- We aim to have a 'good life'.
    - Not just ourselves, but as a society
- Ethically significant harm/benefit happens: "when it has a substantial possibility of making a difference to certain individuals' chances of having a good life, or the chances of a group to live well."
- Ethics implies 'human choice'. Good intentions is not enough to make an ethical choice.
- It is not easy to identify the harms and benefits of data in a specific context. We should increase awareness!

Some Common Ethical Challenges

Ethical Challenges in Appropriate Data Collection and Use

- Purpose of data collection, context, dissemination of data, choice in data sharing, compensation, control/rights...

Data Storage, Security and Responsible Data Stewardship

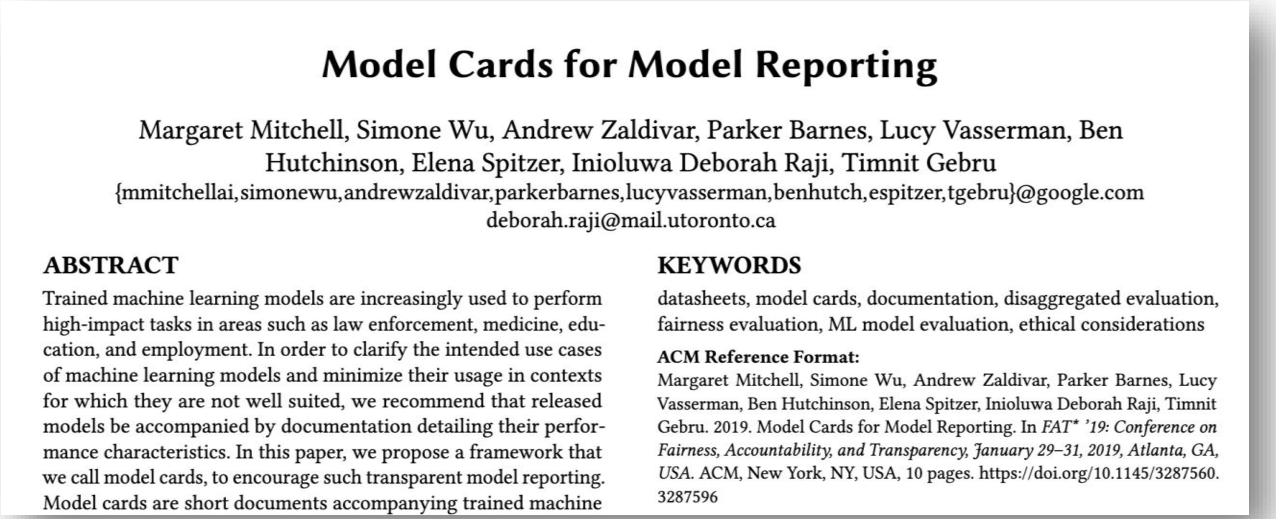
- Storage of data, risk estimation, mitigation strategies, privacy-preserving techniques, ethical risks of keeping data longer...

Datasheets for datasets*

- Bridging the gap between dataset creators and data consumers
- Good for:
  - Reproducibility
  - Increasing accountability and transparency
  - Mitigating unwanted biases
  - Deciding on the use of a dataset
- Similar datasets could be created based on datasheets

*Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (December 2021)

Democratization of ML --- Model Cards



Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

https://modelcards.withgoogle.com/about

Analyzing Case Studies Ethically

12-step Approach Overview

1. State the nature of the ethical issue you've initially spotted

2. List the relevant facts

3. Identify stakeholders

4. Clarify the underlying values

5. Consider consequences

6. Identify relevant rights/duties

7. Reflect on which virtues apply

8. Consider relevant relationships

9. Develop a list of potential responses

10. Use moral imagination to consider each option based on the above considerations

11. Choose the best option

12. Consider what could be done in the future to prevent the problem

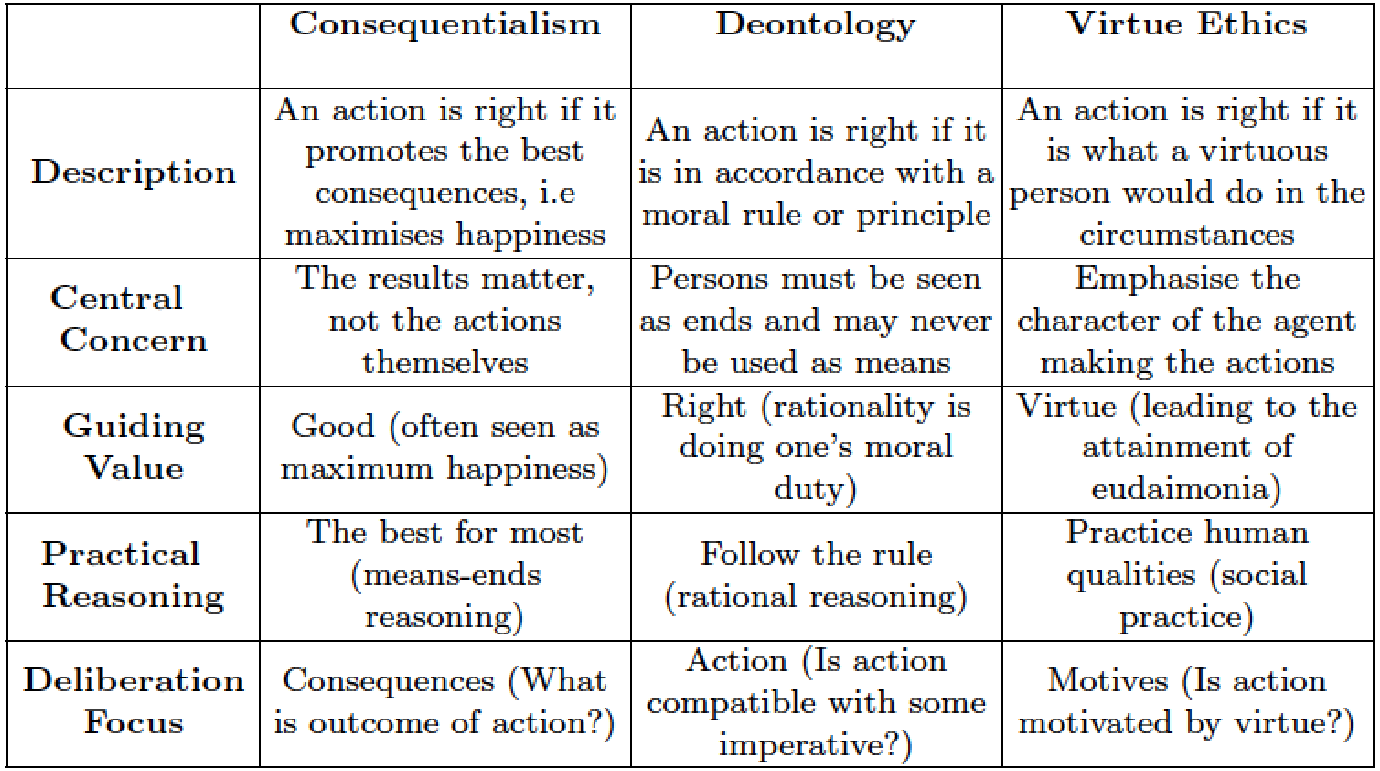
Ethical Decision-Making

MORAL PHILOSOPHY

Ethics/Morality

- We will use these terms interchangeably.
- These terms focus on how humans should act.
- We want to achieve what is right, fair and just, does not cause harm.
- Applicability to various cases is important since philosophers have the tendency to introduce general answers.

Comparison of Main Ethical Theories



	Consequentialism	Deontology	Virtue Ethics
Description	An action is right if it promotes the best consequences, i.e maximises happiness	An action is right if it is in accordance with a moral rule or principle	An action is right if it is what a virtuous person would do in the circumstances
Central Concern	The results matter, not the actions themselves	Persons must be seen as ends and may never be used as means	Emphasise the character of the agent making the actions
Guiding Value	Good (often seen as maximum happiness)	Right (rationality is doing one's moral duty)	Virtue (leading to the attainment of eudaimonia)
Practical Reasoning	The best for most (means-ends reasoning)	Follow the rule (rational reasoning)	Practice human qualities (social practice)
Deliberation Focus	Consequences (What is outcome of action?)	Action (Is action compatible with some imperative?)	Motives (Is action motivated by virtue?)

Dignum, Virginia. "Responsible artificial intelligence: designing AI for human values" (2017)

Making an Ethical Decision

- Markkula Center for Applied Ethics at Santa Clara University has great Ethics Resources.
- You can use 12-step approach by extending the ethical dimension by using the set of ethical questions.
- When it comes to evaluating alternative actions, you can ask the following questions:

Markkula Center
for Applied Ethics
at Santa Clara University

https://www.scu.edu/media/ethics-center/resources/making.pdf

Machine Ethics

Why is it challenging?

How to implement Machine Ethics?

- Top-Down
  - Start with an ethical theory, identify smaller problems and solve them.
  - Pros: no need to identify additional problems
  - Cons: Not clear from the beginning if subproblems are solvable
- Bottom-Up
  - Start with data, and learn ethical behavior from data.
  - Pros: Subproblems are solvable
  - Cons: Non-necessary subproblems may be dealt with.

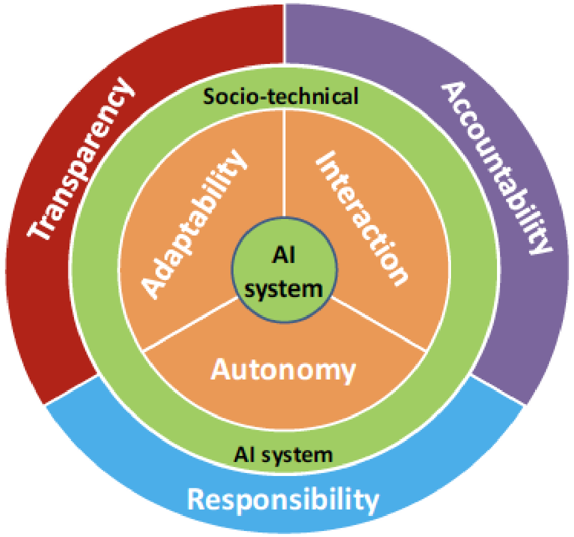
Wallach and Allen. 2008. Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford, UK.

The ART Principles

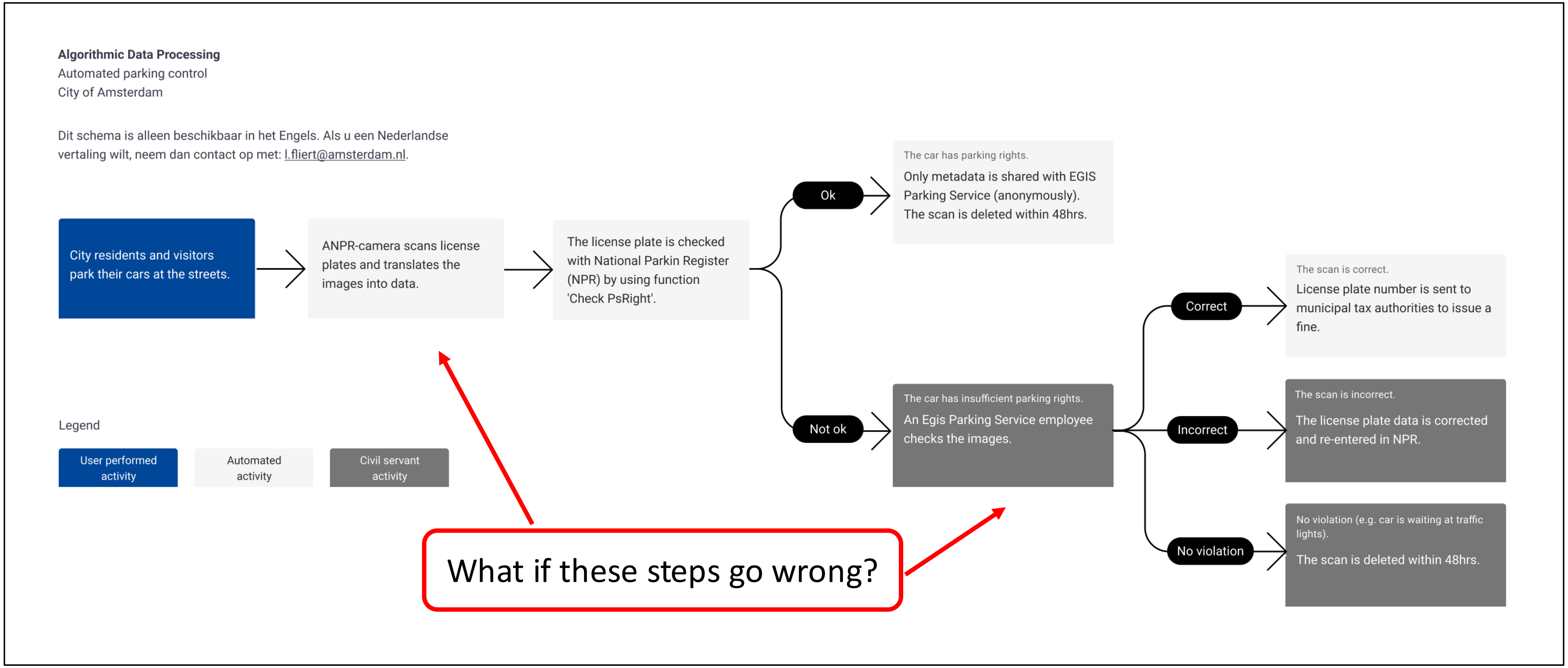
Accountability, Responsibility, Transparency

The ART Principles for Trustworthy Autonomous Systems

- Accountability
  - The system explains and justifies its decision to users and relevant parties.
- Responsibility
  - The focus is on how the socio-technical systems operate.
- Transparency
  - It is about the data being used, methods being applied, openness about choices and decisions.



Transparency: Automated Parking Control



Algorithmic Data Processing
Automated parking control
City of Amsterdam

Dit schema is alleen beschikbaar in het Engels. Als u een Nederlandse vertaling wilt, neem dan contact op met: l.fliert@amsterdam.nl.

City residents and visitors park their cars at the streets.

ANPR-camera scans license plates and translates the images into data.

The license plate is checked with National Parkin Register (NPR) by using function 'Check PsRight'.

Ok

The car has parking rights.
Only metadata is shared with EGIS Parking Service (anonymously). The scan is deleted within 48hrs.

Not ok

The car has insufficient parking rights.
An Egis Parking Service employee checks the images.

Correct

The scan is correct.
License plate number is sent to municipal tax authorities to issue a fine.

Incorrect

The scan is incorrect.
The license plate data is corrected and re-entered in NPR.

No violation

No violation (e.g. car is waiting at traffic lights).
The scan is deleted within 48hrs.

Legend

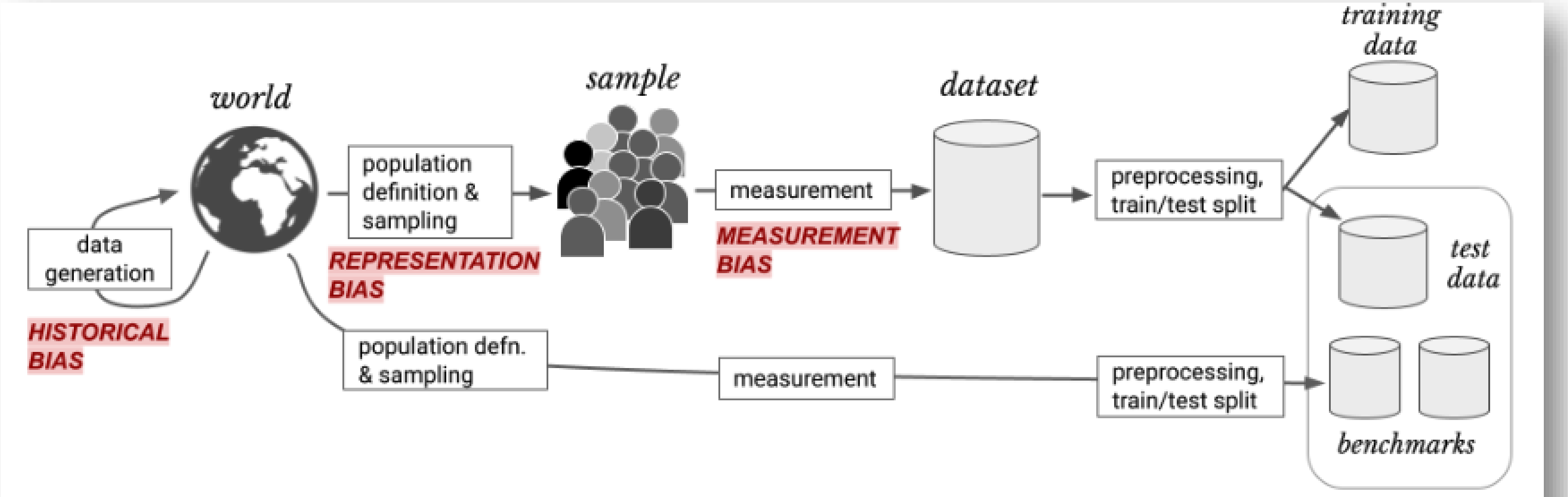
User performed activity	Automated activity	Civil servant activity

What if these steps go wrong?

https://algoritmeregister.amsterdam.nl/en/automated-parking-control/

Justice, Fairness, Bias

The Big Three

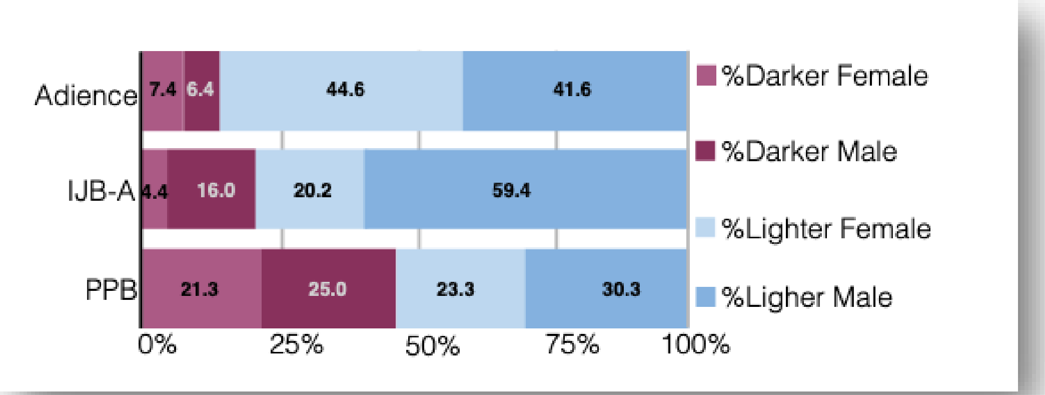


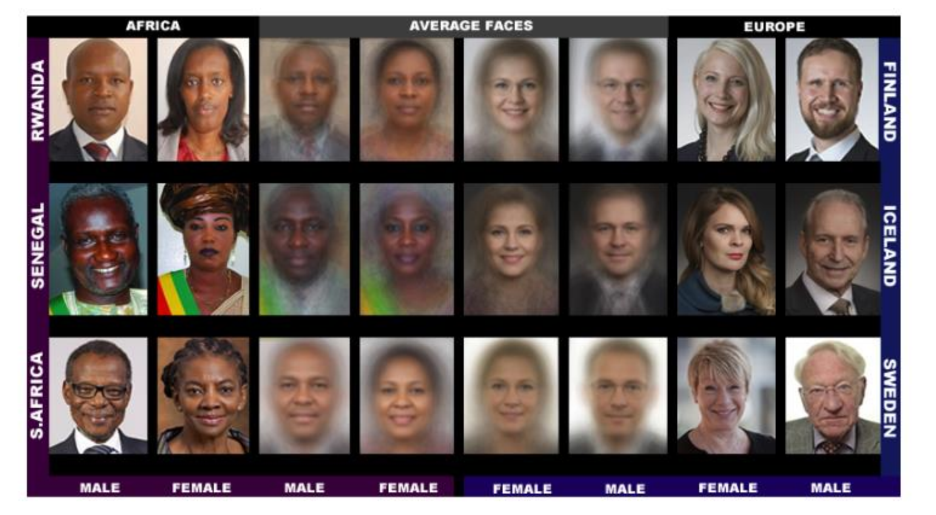
(a) Data Generation

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9.

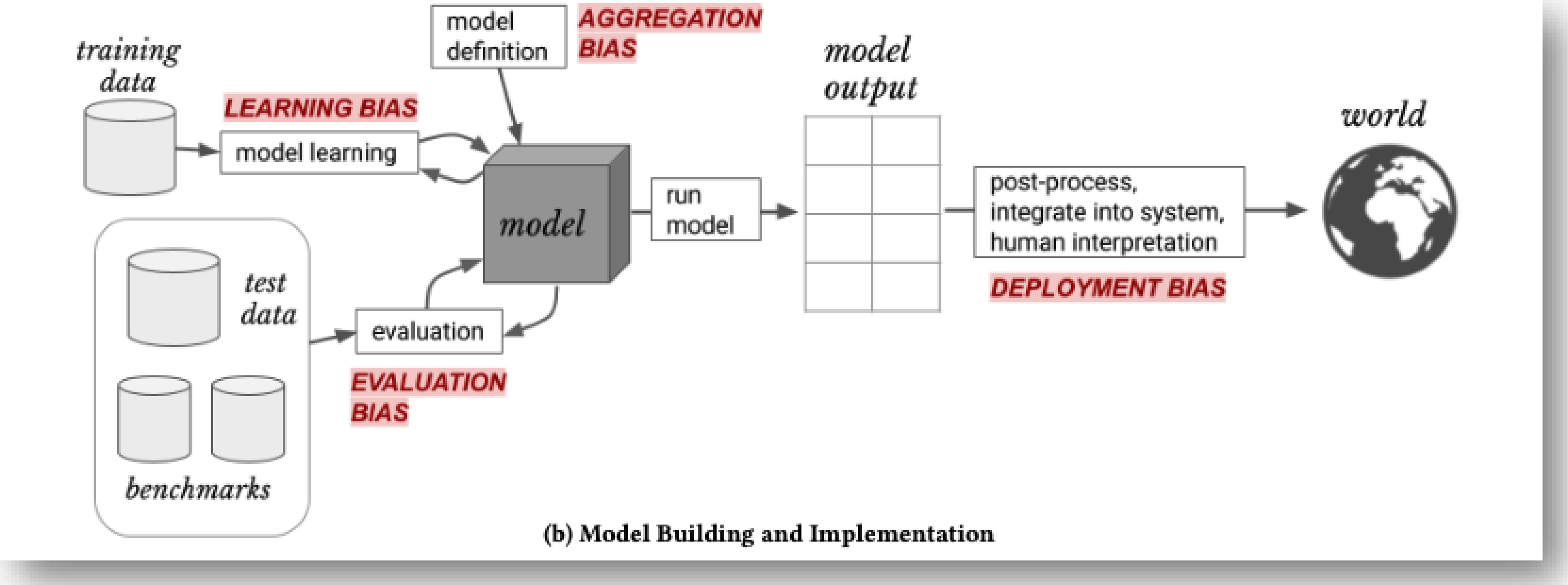
Gender Shades

- Buolamwini and Gebru analyze two benchmarks to report gender and skin type distribution.





Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77-91. PMLR, 2018.



(b) Model Building and Implementation

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9.

Algorithmic Fairness

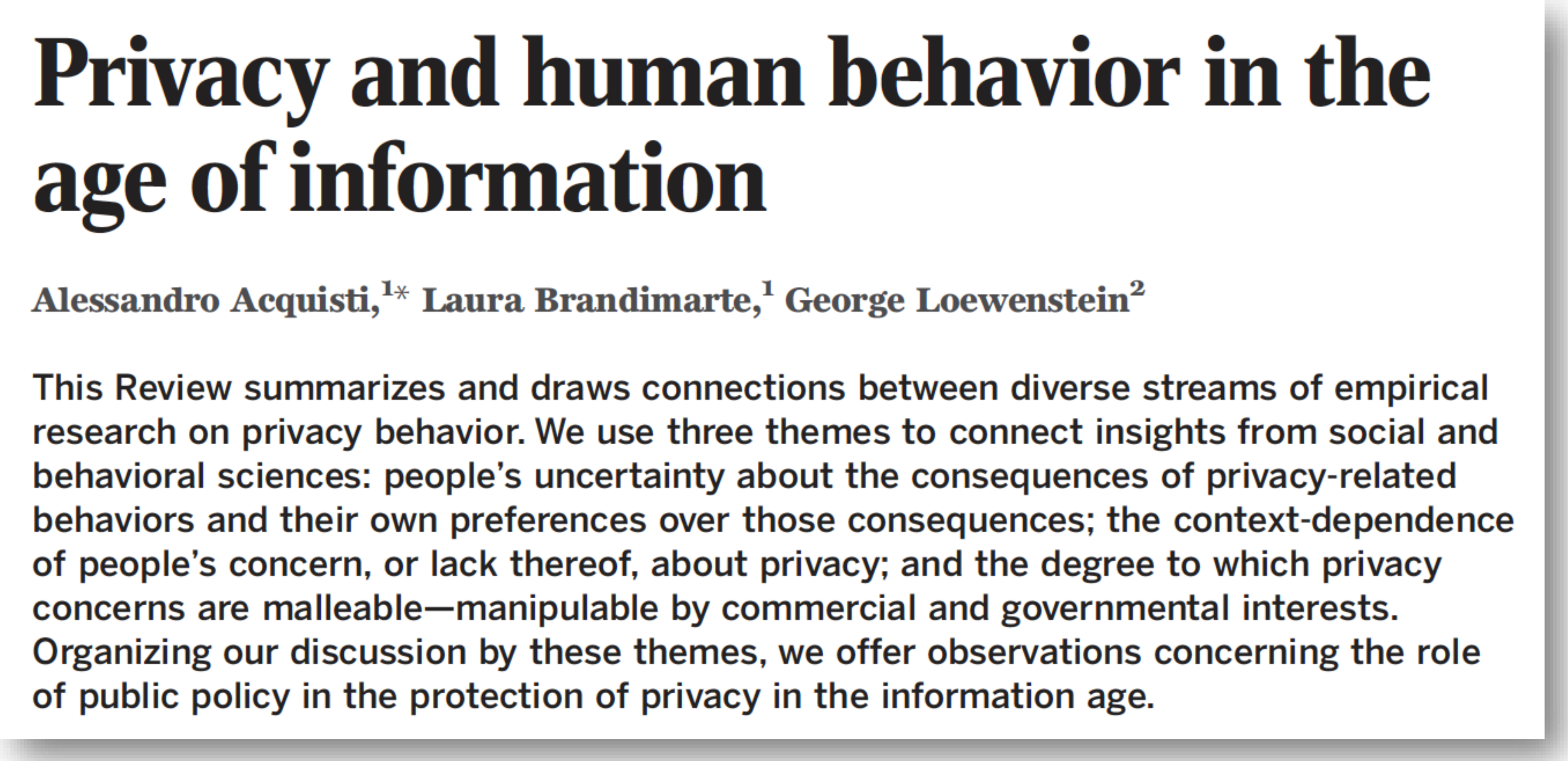
- We can talk about fairness when people are not discriminated against based on their membership to a specific group.
- Fairness definition? The most famous discussion about fairness definitions come from Arvind Narayanan.
- There are two main categories: group fairness (statistical fairness) and individual fairness.

21 Definitions of Fairness -- https://www.youtube.com/watch?v=jIXIuYdnyyk

Algorithmic Justice League (AJL)

- The Algorithmic Justice League is an organization that combines art, research, policy guidance and media advocacy to illuminate the social implications and harms of AI.
- AJL is a cultural movement towards
  - Equitable AI (agency and control, affirmative consent, centering justice)
  - Accountable AI (transparency, continuous oversight, redress harms)
- AJL recognizes the limitations of Ethical AI, which does not create any mandatory requirements or ban certain uses of AI. They focus on creating action.

Privacy and Surveillance

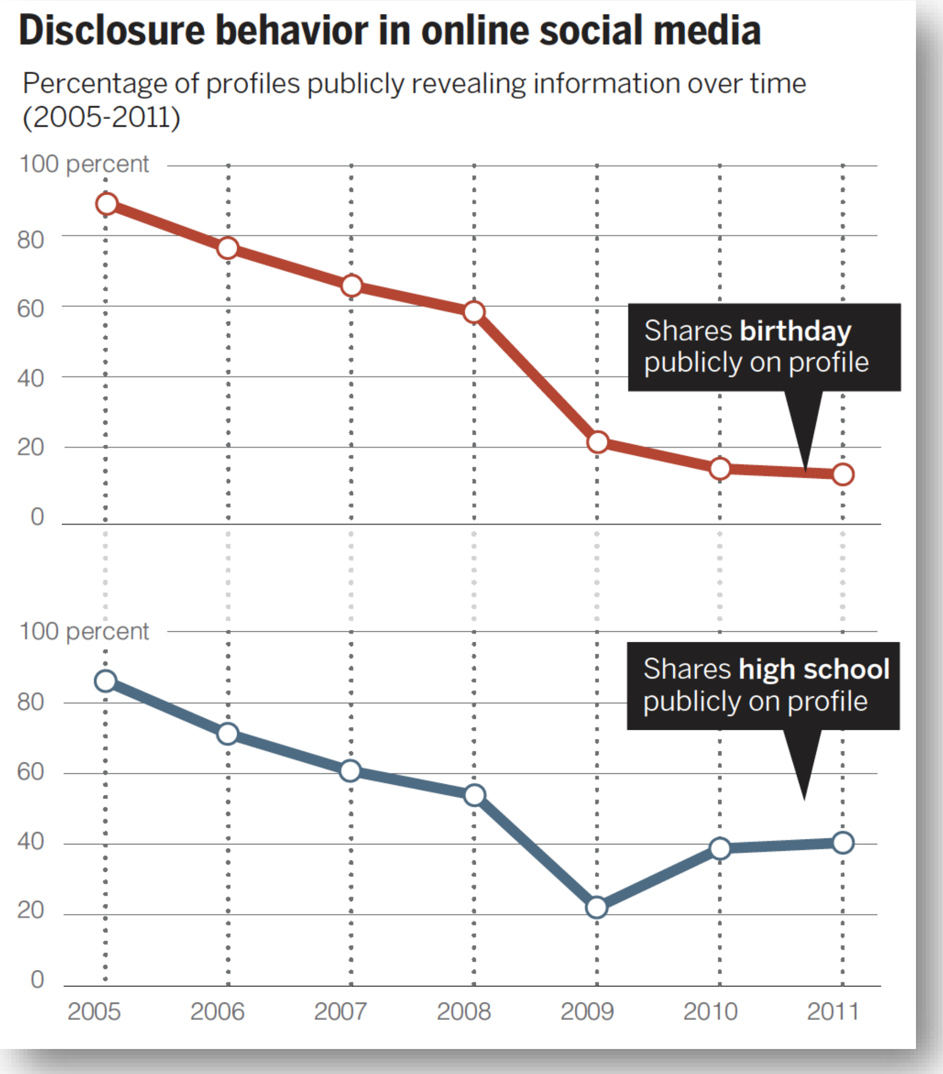


Privacy and human behavior in the age of information

Alessandro Acquisti,[1]* Laura Brandimarte,[1] George Loewenstein[2]

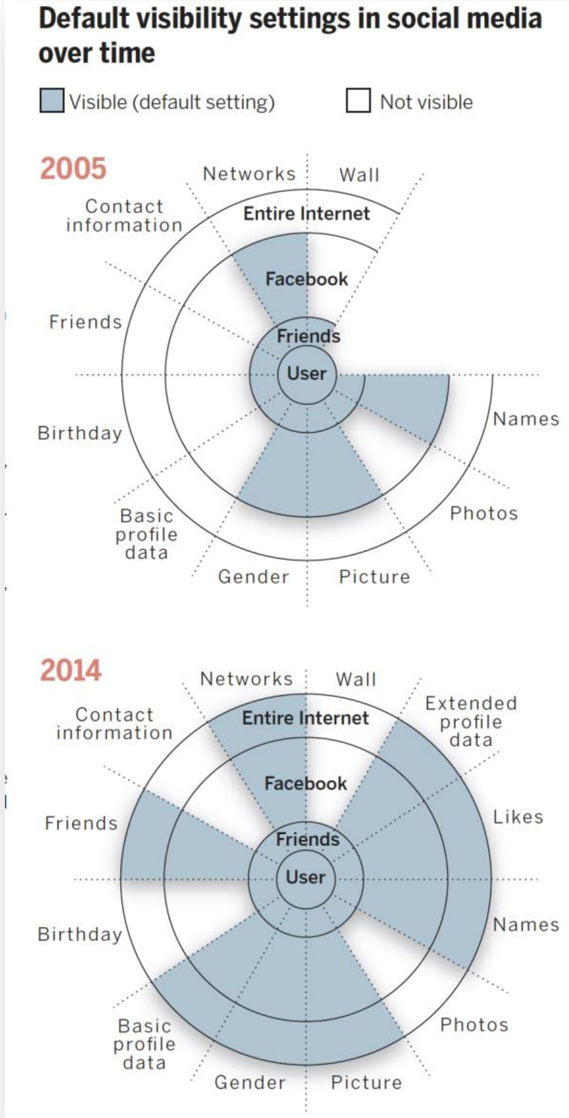
This Review summarizes and draws connections between diverse streams of empirical research on privacy behavior. We use three themes to connect insights from social and behavioral sciences: people's uncertainty about the consequences of privacy-related behaviors and their own preferences over those consequences; the context-dependence of people's concern, or lack thereof, about privacy; and the degree to which privacy concerns are malleable—manipulable by commercial and governmental interests. Organizing our discussion by these themes, we offer observations concerning the role of public policy in the protection of privacy in the information age.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509-514.

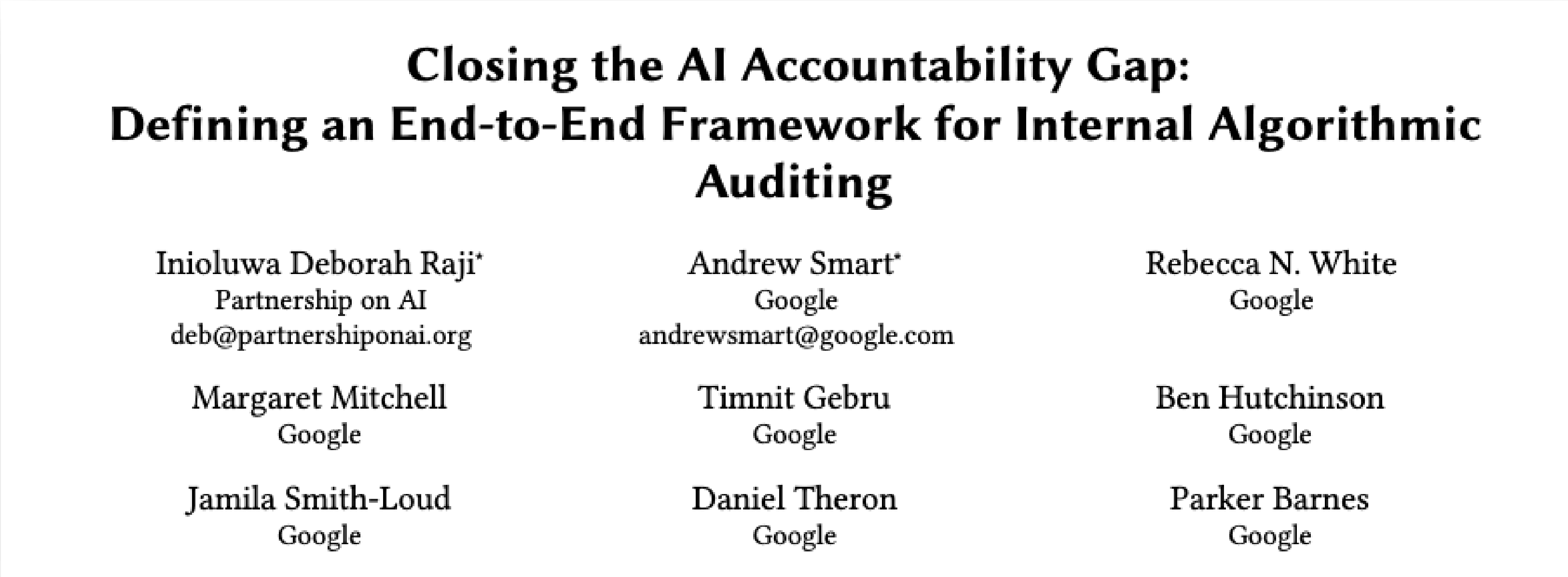


Disclosure behavior in online social media

Percentage of profiles publicly revealing information over time (2005-2011)



Default visibility settings in social media over time



Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Inioluwa Deborah Raji[*]
Partnership on AI
deb@partnershiponai.org

Andrew Smart[*]
Google
andrewsmart@google.com

Rebecca N. White
Google

Margaret Mitchell
Google

Timnit Gebru
Google

Ben Hutchinson
Google

Jamila Smith-Loud
Google

Daniel Theron
Google

Parker Barnes
Google

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20). Association for Computing Machinery, New York, NY, USA, 33–44.*

Guidance Outline

- ICO is focusing on a risk-based approach to AI
  - Assessing the risks to the rights and freedoms of individuals that may arise
- Guidance prepared for:
  - an audience with a compliance focus (e.g., data protection officers (DPOs), ICO’s own auditors)
  - technology specialists (e.g., developers)
- Four main topics are covered:
  - Accountability and governance of AI
  - Fair, lawful, transparent processing
  - Data minimisation and security
  - Rights in AI systems
- Controls: Preventative, Detective, Corrective

THE END