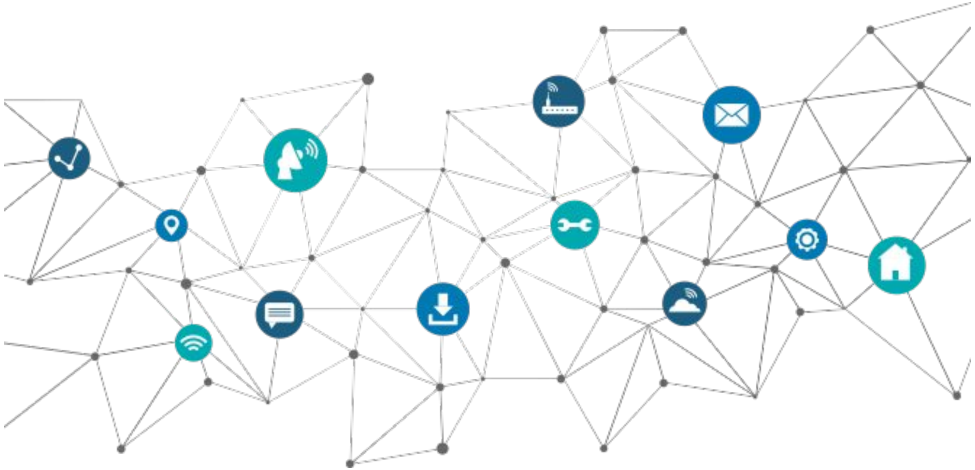
Social and Ethical Issues in AI

Nadin KOKCIYAN

School of Informatics,
University of Edinburgh, UK
nadin.kokciyan@ed.ac.uk



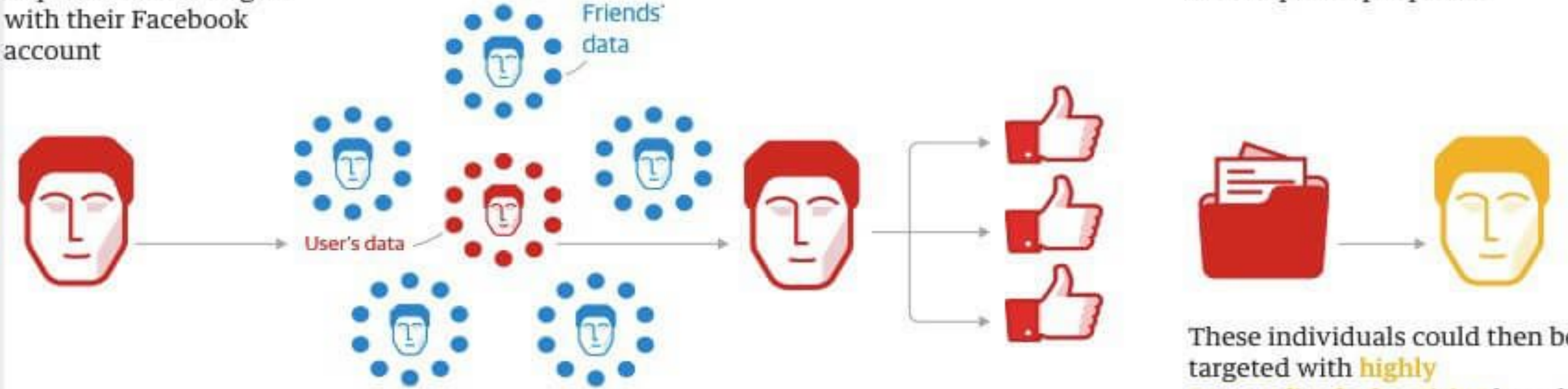
Cambridge Analytica: how 50m Facebook records were hijacked

1
Approx. 320,000 US voters ('seeders') were **paid $2-5 to take a detailed personality/ political test** that required them to log in with their Facebook account

2
The app also **collected data such as likes and personal information** from the test-taker's Facebook account ...

3
The **personality quiz results** were paired with their Facebook data - such as **likes** - to seek out psychological patterns

4
Algorithms combined the data with other sources such as voter records to **create a superior set of records (initially 2m people in 11 key states*)**, with hundreds of data points per person



... as well their **friends'** data, amounting to over 50m people's raw Facebook data

These individuals could then be targeted with **highly personalised advertising** based on their personality data

Guardian graphic. *Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

Amazon Rekognition

Automate your image and video analysis with machine learning.

Get Started with Amazon Rekognition

Amazon Rekognition makes it easy to add image and video analysis to your applications using proven, highly scalable, deep learning technology that requires no machine learning expertise to use. With Amazon Rekognition, you can identify objects, people, text, scenes, and activities in images and videos, as well as detect any inappropriate content. Amazon Rekognition also provides highly accurate facial analysis and facial search capabilities that you can use to detect, analyze, and compare faces for a wide variety of user verification, people counting, and public safety use cases.

BBC PANORAMA

Amazon: What They Know About Us
https://www.bbc.co.uk/programmes/m000fjz

Facial recognition use by South Wales Police ruled unlawful

By Jenny Rees
BBC Wales home affairs correspondent

11 August 2020

"For three years now, South Wales Police has been using it against hundreds of thousands of us, without our consent and often without our knowledge."

"We should all be able to use our public spaces without being subjected to oppressive surveillance."

https://www.bbc.co.uk/news/uk-wales-53734716

8 Ethical Questions in AI

Bias:
Is AI fair?

Liability:
Who is responsible for AI?

Security:
How do we protect access to AI from bad actors?

Human Interaction:
Will we stop talking to one another?

Employment:
Is AI getting rid of jobs?

Wealth Inequality:
Who benefits from AI?

Power & Control:
Who decides how to deploy AI?

Robot Rights:
Can AI suffer?

www.logikk.com

LOGIKK

What is 'Ethics'?

- "Ethics is concerned with studying and/or building up a coherent set of rules or principles by which people ought to live".
- We all have some 'rules of thumbs' that define our behavior.
    - It is right to …
    - It is wrong to …

Let's start with a 'simple' rule

- It is wrong to kill.
  - Is it wrong to kill animals?
  - Is killing in self-defense wrong?
  - Is the termination of pregnancy wrong?
  - ...

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.

Some Ethical Theories

- Virtue Theories:
  - Who is doing the action?
- Consequentialist Theories:
  - Are the consequences moral?
- Deontological Theories:
  - Is the action itself moral?

	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?)

Machine Ethics

Humans are machines and humans have ethics.

Machine ethics does not exist because ethics is simply emotional

Can a computer operate ethically because it is internally ethical in some way?

The Nature, Importance, and Difficulty of Machine Ethics | *James H. Moore, IEEE Intelligent Systems, 2006*

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 behaviour 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. DOI:10.1093/acprof:oso/9780195374049.001.0001

The Dilemma of a Rescue Robot

A recent experiment conducted by Alan Winfield and colleagues shows that rescue robots may enter into ethical dilemmas, see [1]. In the experiment, A (for Asimov), a robot, is saving (robot stand-ins for) human beings who are about to move into a dangerous area. This the robot does by moving in front of them, which causes them small discomfort but also has the effect that they turn away from danger. However, in case of exact symmetry in terms of distance between the human beings to be saved, the robot may dither between saving one or the other and thus fail to save anyone.

[1] Alan FT Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M.Witkowski, and C. Melhuish, editors, Advances in Autonomous Robotics Systems, pages 85–96. Springer, 2014.

Specification in YAML

Software used:
http://www.hera-project.com/

rescue-robot.yaml — examples (git: master)

```yaml
description: The Rescue Robot Dilemma
actions: [a_save_h1, a_save_h2, a_remain_inactive]
background: [b_save_people]
consequences: [saved_h1, discomfort_h1, saved_h2, discomfort_h2]
mechanisms:
    saved_h1: And("b_save_people", "a_save_h1")
    discomfort_h1: a_save_h1
    saved_h2: And("b_save_people", "a_save_h2")
    discomfort_h2: a_save_h2
utilities:
    saved_h1: 10
    discomfort_h1: -4
    saved_h2: 10
    discomfort_h2: -4
    Not('saved_h1'): -10
    Not('discomfort_h1'): 4
    Not('saved_h2'): -10
    Not('discomfort_h2'): 4
intentions:
    a_save_h1: [a_save_h1, saved_h1]
    a_save_h2: [a_save_h2, saved_h2]
    a_remain_inactive: [a_remain_inactive]
```

The ART Principles

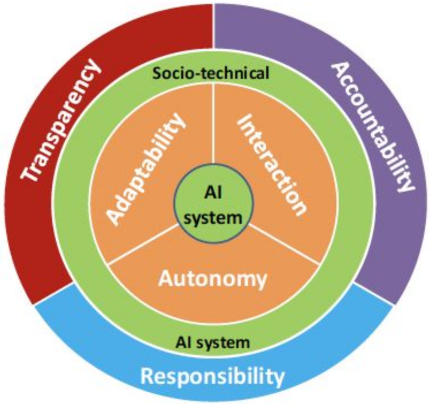
Accountability, Responsibility, Transparency

AI has great potential (if controlled)

- AI can bring significant benefits to society.
  - e.g., climate change, cure to diseases …
- As we mentioned so far in the lectures, AI can produce undesirable impacts.
  - e.g., amplifying biases, discrimination, misinformation, manipulation …
- We need to find an ethically acceptable way of designing technology that can benefit the society.

The ART Principles for Trustworthy Autonomous Systems

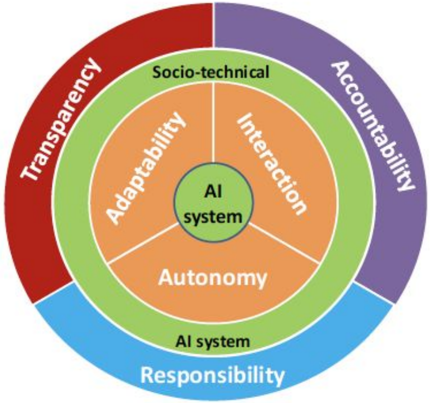
- Accountability
  - The system explains and justifies its decision to users and relevant parties.
- Responsibility
  - The focus is on how the socio-technical systems operate.
- Transparency
  - It is about the data being used, methods being applied, openness about choices and decisions.



The ART Principles for Trusty tonomous Systems

- Accountabilit
  - The s
    re
- Res
  - T al systems operate.
  - It olders and the machines.
- Transpa
  - It is a being used, methods being applied,
    openn choices and decisions.

ART is essential to build social trust in Autonomous Systems



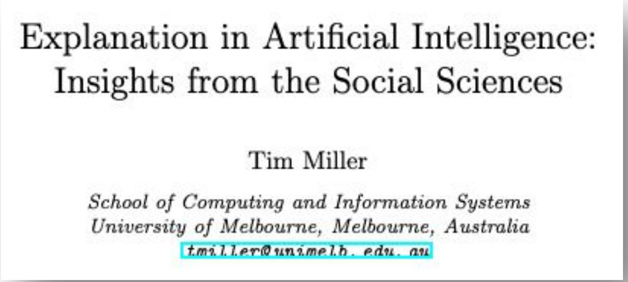
Transparency

- Many other terms: "explainability", "understandability", "interpretability"
- Transparency in AI:
  - supports access to justifications for decisions when needed. In public sector, people should also know how to contest and appeal.
  - addresses the right to know (e.g., GDPR). For example, a participation information sheet should include all details about data lifecycle.
  - helps in understanding and managing risks. For example, an organization can be responsible and accountable if it knows the inner workings of their offered solutions.

Major Findings from the literature on explanations

According to Miller, explanations are:

- Contrastive
  "Why event P happened instead of some event Q?"
- Selected (influenced by cognitive biases)
  (Partial) explanations are based on <u>selected</u> factors
- Not driven by probabilities
  Effective explanations are causal, not the most likely explanations
- Social/interactive
  Explanations <u>for</u> the user

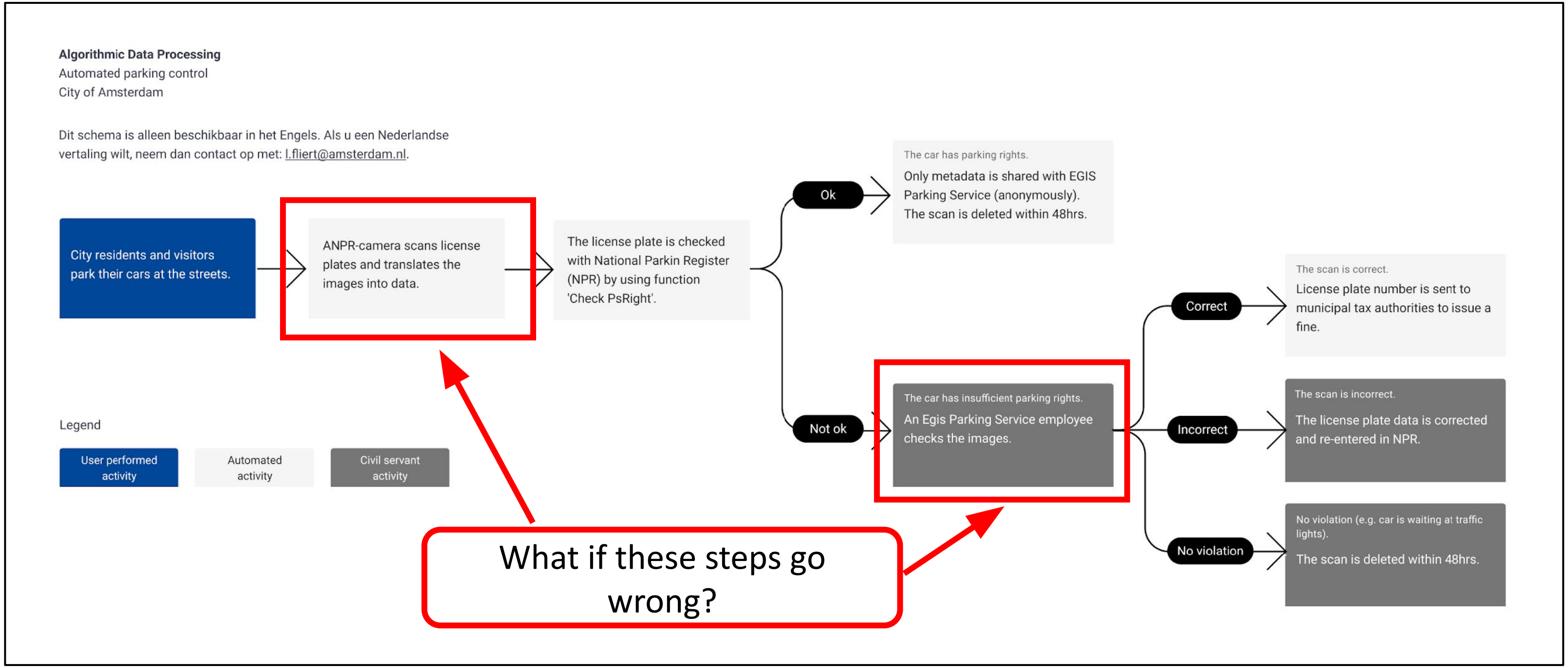

Explanation in Artificial Intelligence: Insights from the Social Sciences

Tim Miller

School of Computing and Information Systems
University of Melbourne, Melbourne, Australia
tmiller@unimelb.edu.au

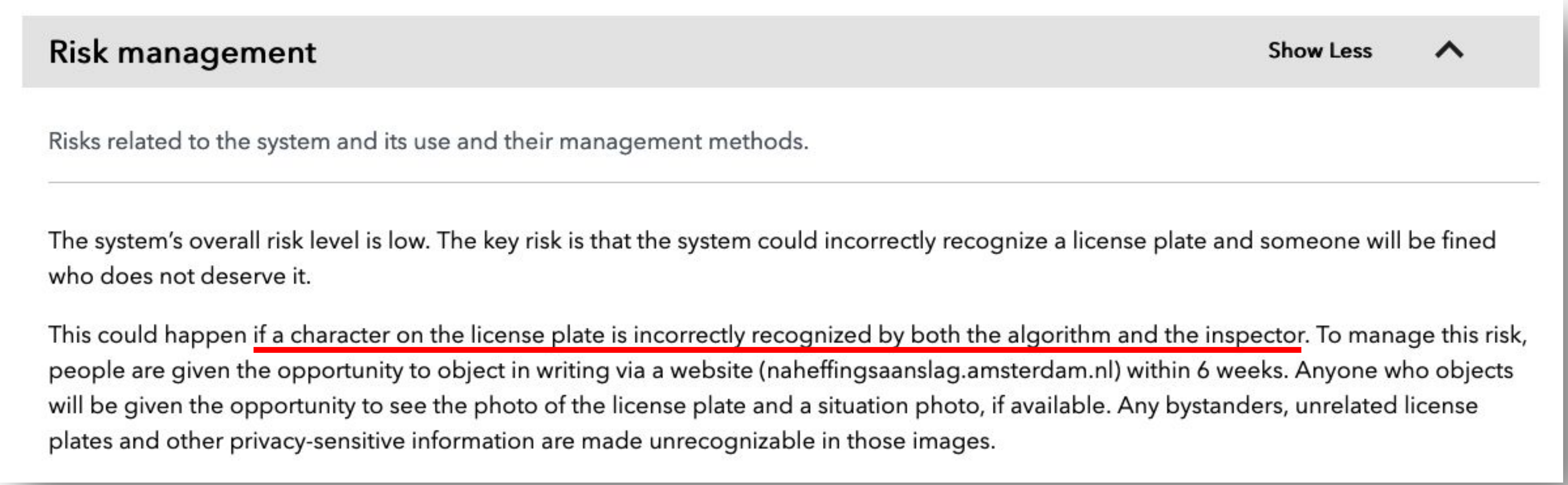
Miller, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267 (2019), 1-38.

Transparency: Automated Parking Control



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

Transparency: Automated Parking Control


Risk management

Show Less

Risks related to the system and its use and their management methods.

The system's overall risk level is low. The key risk is that the system could incorrectly recognize a license plate and someone will be fined who does not deserve it.

This could happen if a character on the license plate is incorrectly recognized by both the algorithm and the inspector. To manage this risk, people are given the opportunity to object in writing via a website (naheffingsaanslag.amsterdam.nl) within 6 weeks. Anyone who objects will be given the opportunity to see the photo of the license plate and a situation photo, if available. Any bystanders, unrelated license plates and other privacy-sensitive information are made unrecognizable in those images.

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

Transparency: Automated Parking Control

Data processing

Show Less

The operational logic of the automatic data processing and reasoning performed by the system and the models used.

Model architecture

The service uses license plate recognition algorithms to locate and process the license plate data from the camera data stream. Algorithms are used to locate the license plate from the image data, to adjust the images for identification, to identify the individual characters of the license plate, and to validate the plate contents against national license plate characteristics.

After a successful plate identification and processing, license plate data is sent to the National Parking Register for further processing. NPR's algorithm checks the validity of parking rights for the license plate in a given time and location (for technical information on the NPR algorithm, see the information on their website: https://nationaalparkeerregister.nl/fileadmin/files/Mobiel_parkeren /Interface_Description_v7.6.pdf). A positive response means the car has valid parking rights in place, and the license plate scan data can be removed in 48 hours. For license plates with invalid parking rights, the case is transferred to the cities tax department, which connects to the RDW database to link the license plate with the car ownership data, and to deliver a fine.

Content	Attachment
System architecture description	Automated parking control Attach architecture image

They provide 58 pages to explain the algorithm!

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

Why is transparency hard?

- We are talking about sociotechnical systems; hence we are dealing with many stakeholders.
- Contexts, user profiles, questions to be answered vary largely.
- A data scientist may need to learn more about unjust biases in their data, whereas a user may be interested in something different.

Why is transparency hard?

- How to explain the workings of a "black box" model?
  - Explanations could be added by design, but this requires careful engineering to have a usable solution (e.g., interactive interfaces are great to explore models)
  - The use of simpler models works sometimes!
- How much transparency should we provide? We do not want to make our systems vulnerable to attacks at the same time.

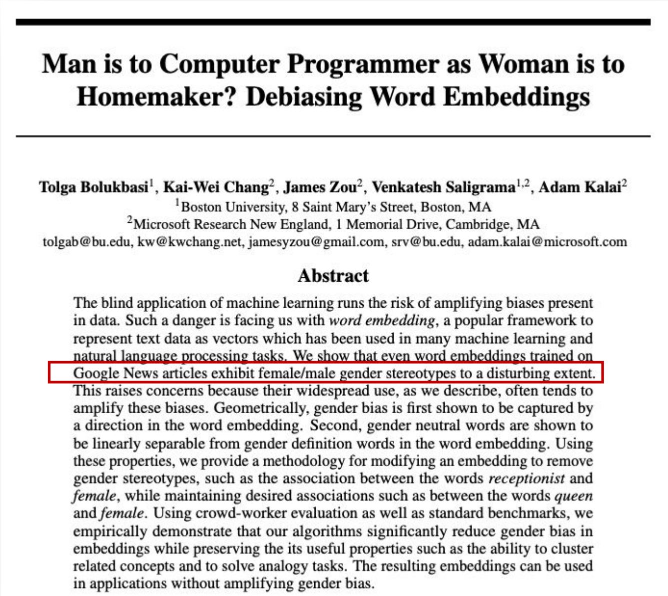
Justice, Fairness, Bias

The Big Three

Justice, Fairness and Bias

- Kant emphasizes the importance of human dignity.
- Individuals expect to be treated fairly; the violation of human dignity leads to discrimination.
- Discrimination is the unjust treatment of people based on the groups or classes they belong to. Discrimination may stem from biases.
- We often talk about algorithmic fairness, since algorithms may amplify existing economic and societal bias.

Discrimination and Biases

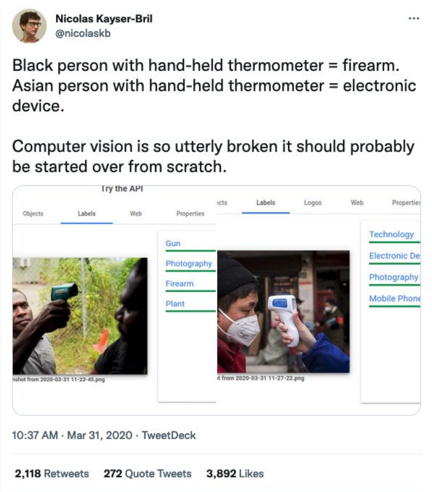


Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.



A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Harini Suresh
John Guttag
hsuresh@mit.edu
guttag@mit.edu

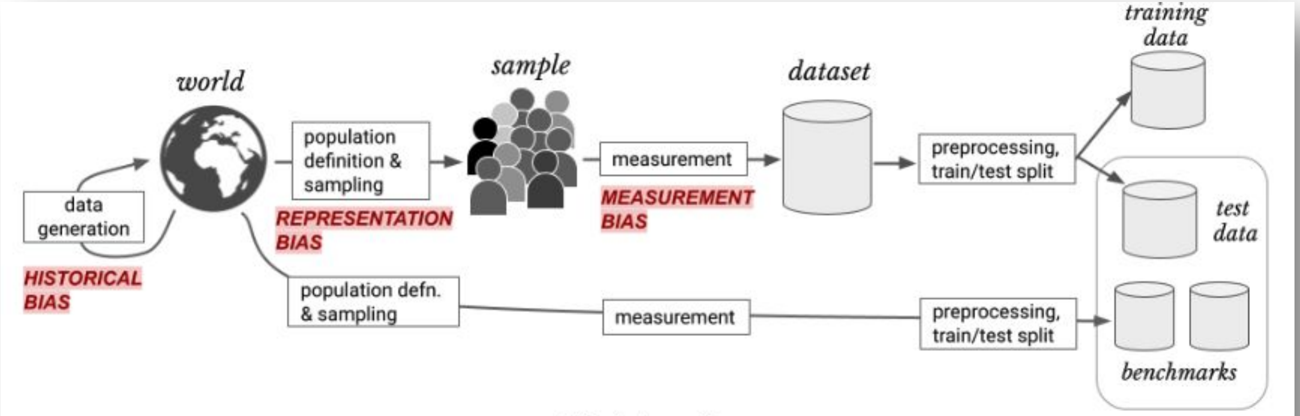
ABSTRACT

As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle. In this paper, we provide a framework that identifies seven distinct potential sources of downstream harm in machine learning, spanning data collection, development, and deployment. In doing so, we aim to facilitate more productive and precise communication around these issues, as well as more direct, application-grounded ways to mitigate them.

necessarily because the statement “data is biased” is *false*, but because it treats data as a static artifact divorced from the process that produced it. This process is long and complex, grounded in historical context and driven by human choices and norms. Understanding the implications of each stage in the data generation process can reveal more direct and meaningful ways to prevent or address harmful downstream consequences that overly broad terms like “biased data” can mask.

Moreover, 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

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.



(a) Data Generation

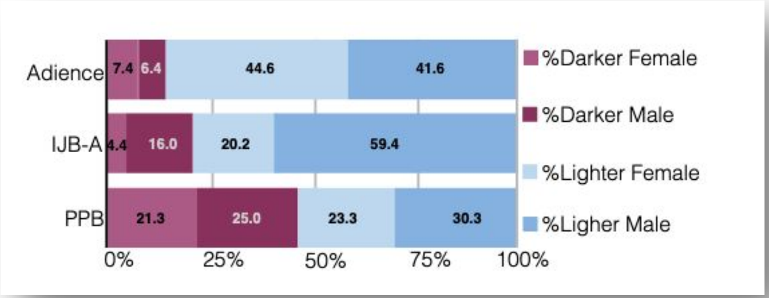
Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9.

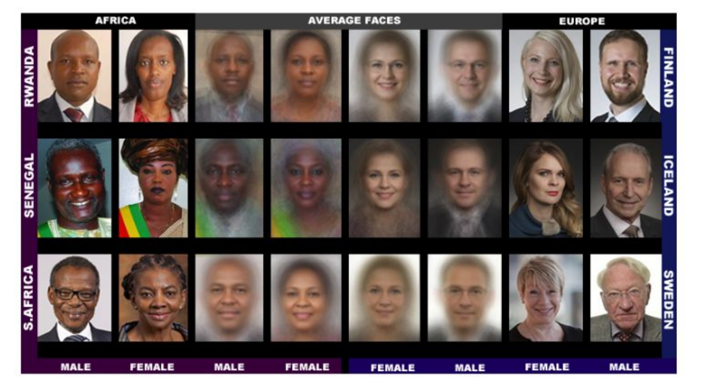
Representation Bias

- Target population does not reflect the use population
  - Model is trained on population X and applied to population Y
  - Model is trained on the same population in different time frames
- Target population contains under-represented groups
  - For example, some age groups may not be represented well in the data
- Sampling method is limited (sampling bias)
  - Target population is set to X, but the data available is only a small subset of X

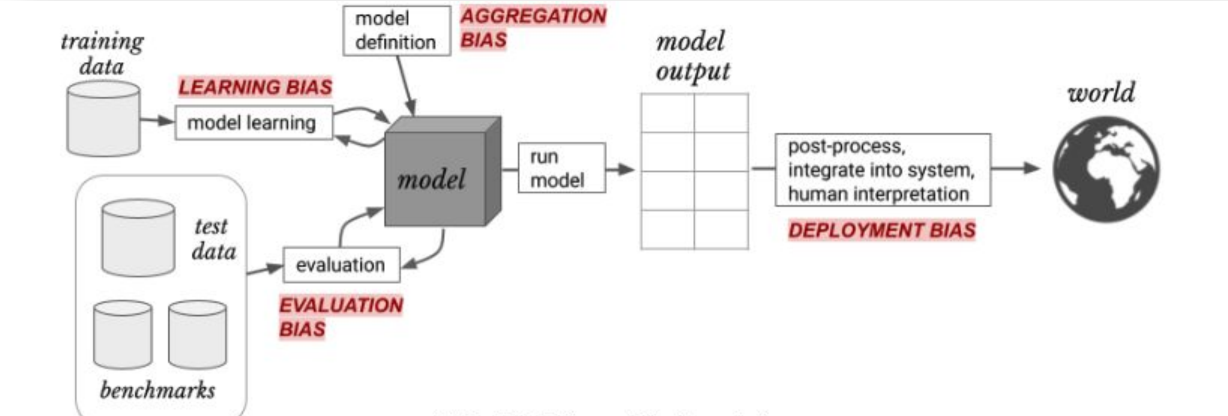
Gender Shades

- 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.

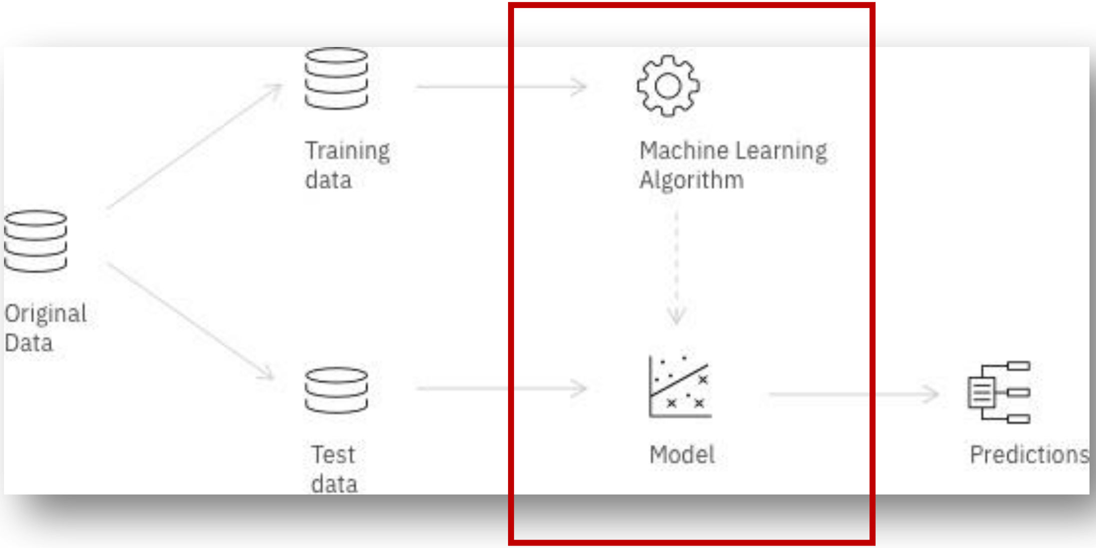


(b) Model Building and Implementation

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.

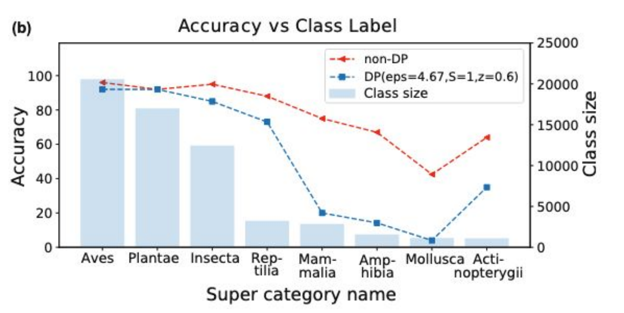
Learning Bias

- Learning bias happens when modeling choices amplify performance disparities.



Disparate Impact on Model Accuracy

- Differential privacy (DP) comes with a cost, which is a reduction In the model's accuracy.
- Bagdasaryan *et al.* show that accuracy of models, trained with DP stochastic gradient descent, drops much more for the underrepresented classes and subgroups.
- This gap is bigger in the DP model than in the non-DP model.
- The results are reported from the sentiment analysis of text and image classification.



Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov. "Differential privacy has disparate impact on model accuracy." *Advances in Neural Information Processing Systems* 32 (2019).

Evaluation Bias

- Evaluation bias occurs when the benchmark datasets (e.g., ImageNet) do not represent the use population.
- The choice of metrics can also result in evaluation bias (e.g., aggregate results, reporting one type of metric)

Gender Shades (Evaluation/Learning Bias example)

- They use their dataset (PPB) to evaluate three commercial gender classification systems (Microsoft, IBM, Face++):

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

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.

De-biasing Algorithms

- Increasing awareness about different types of bias is essential.
- We will now have a closer look at how to design an AI system that would not discriminate.

Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zemel¶

November 30, 2011

Abstract

We study *fairness in classification*, where individuals are classified, e.g., admitted to a university, and the goal is to prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier (the university). The main conceptual contribution of this paper is a framework for fair classification comprising (1) a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the

2018 ACM/IEEE International Workshop on Software Fairness

Fairness Definitions Explained

Sahil Verma
Indian Institute of Technology Kanpur, India
vsahil@iitk.ac.in

Julia Rubin
University of British Columbia, Canada
mjulia@ece.ubc.ca

ABSTRACT

Algorithm fairness has started to attract the attention of researchers in AI, Software Engineering and Law communities, with more than twenty different notions of fairness proposed in the last few years. Yet, there is no clear agreement on which definition to apply in each situation. Moreover, the detailed differences between multiple definitions are difficult to grasp. To address this issue, this paper

training data containing observations whose categories are known. We collect and clarify most prominent fairness definitions for classification used in the literature, illustrating them on a common, unifying example – the German Credit Dataset [18]. This dataset is commonly used in fairness literature. It contains information about 1000 loan applicants and includes 20 attributes describing each applicant, e.g., credit history, purpose of the loan, loan amount

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

Thank you!

Nadin Kokciyan

nadin.kokciyan@ed.ac.uk

https://homepages.inf.ed.ac.uk/nkokciya/

https://twitter.com/nkokciyan