2025/26

Case Studies in AI Ethics (CSAI)

CSAI Teaching Team

- Nadin Kokciyan [Weeks 1-5], Daniel Woods [Weeks 6-10] (Lecturers)
- Rhodri Thomas (Teaching Assistant)
- Nadin Kokciyan (Tutor)
- Sydelle de Souza (Marker)
- Ivan Vegner (Marker)

Course Structure

- Lectures (@ AT Lecture Theatre 3) – Mondays 3:10-5pm
- Discussions (on Learn) – Thursdays
- Questions (on Piazza)

- Tutorials (Week 4 and Week 7)
- Courseworks (CW1 [0%], CW2 [40%])
- Exam [60%]

Courseworks

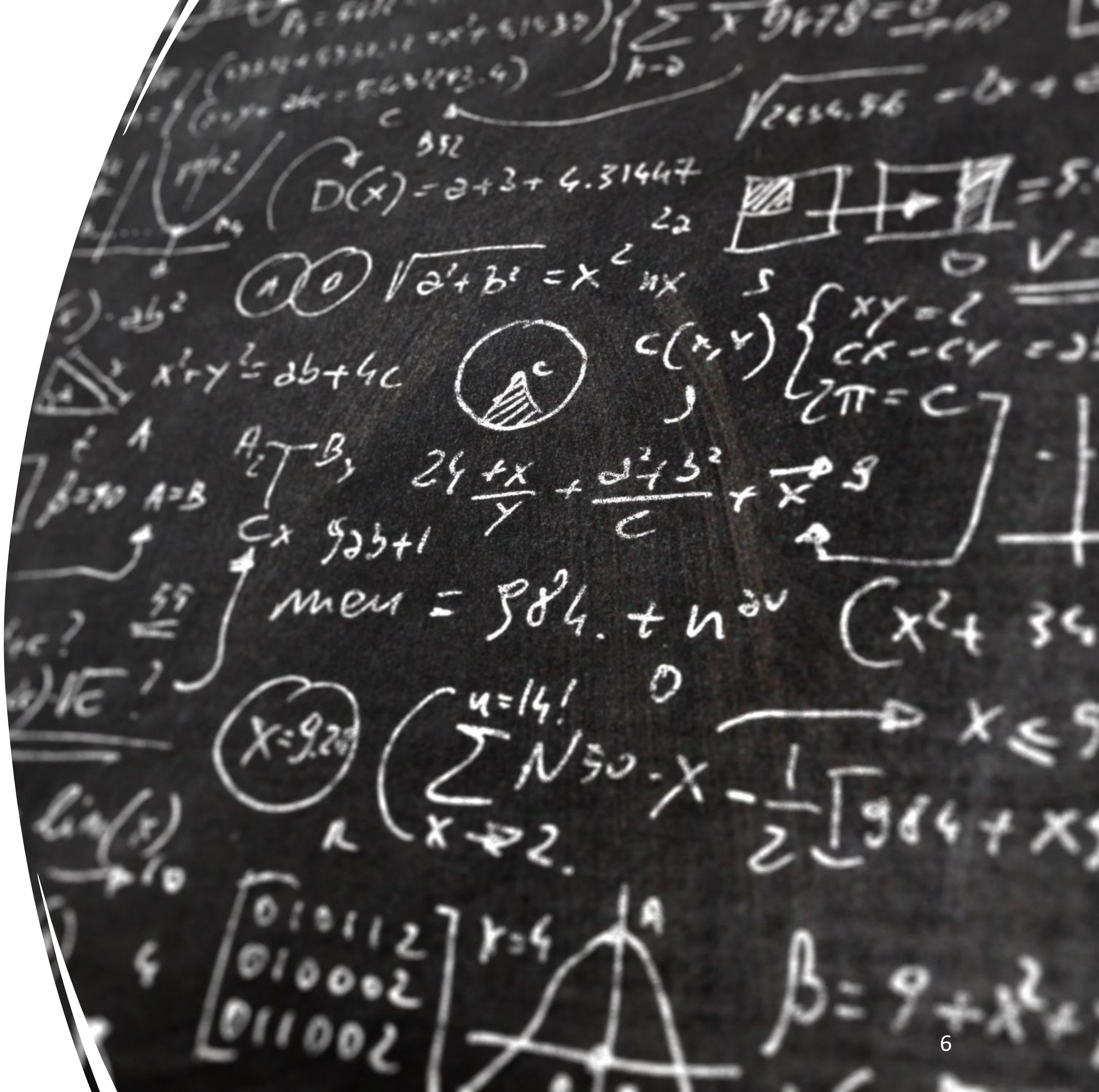
- **CW1: Design Outline (0%) -** This is a group coursework. Each group will select a case study, the students will then **provide an outline** detailing ethical issues that they would like to work on during CW2.
- **CW2: Essay (40%) -** Each student will **write an essay** based on the outline submitted as CW1.

Assignment	Released	Submit by	Feedback by
CW1 Essay Outline (Group)	26/01/2026	Monday 09/02/2026 12:00	23/02/2026
CW2 Essay (Individual)	23/02/2026	Monday 16/03/2026 12:00	06/04/2026

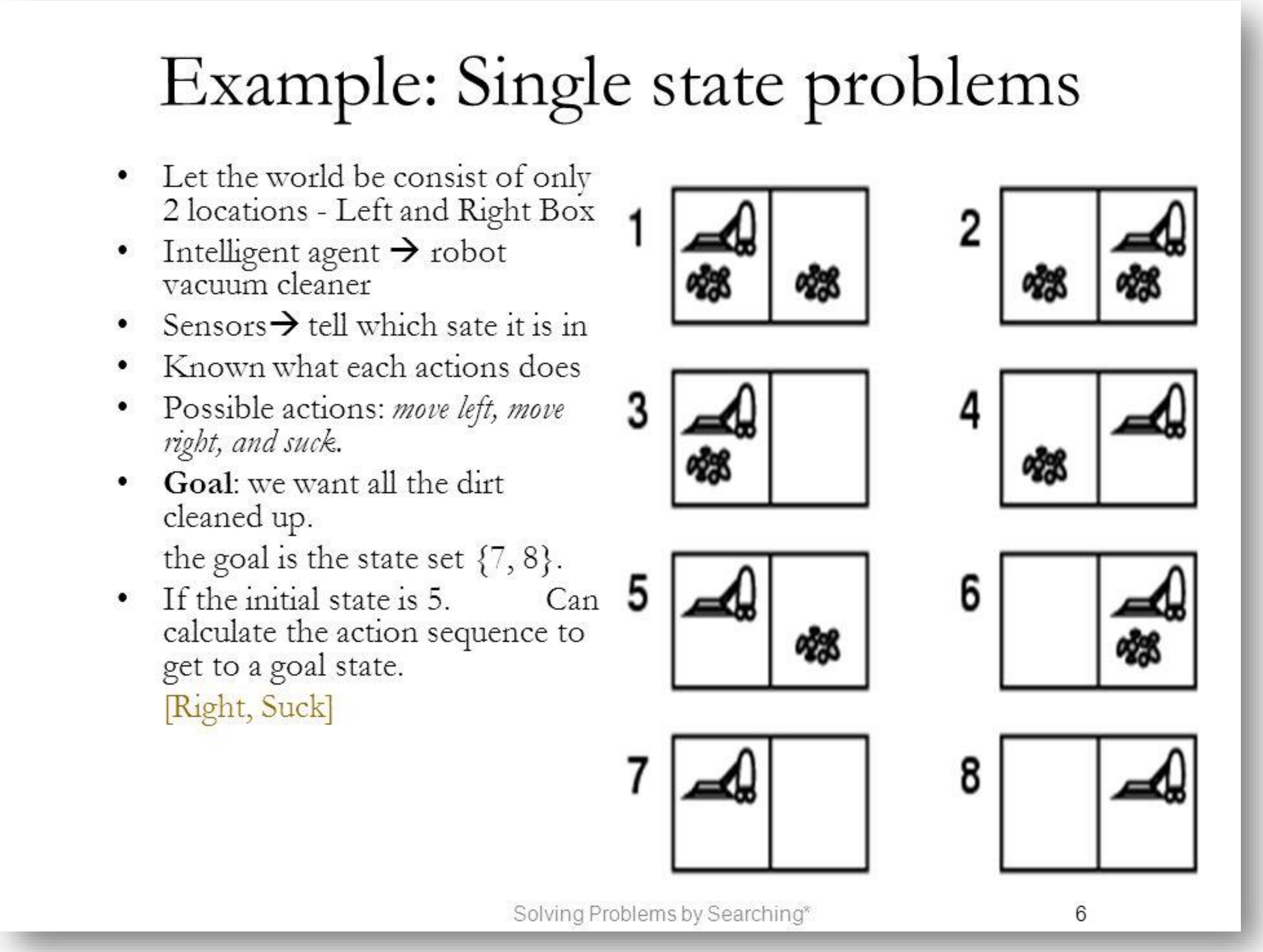
AI, Technology and Ethics

AI?

- Bellman (1978) defines AI as "the automation of activities that we associate with human thinking (i.e., cognitive activities)".
- Hence, the focus is on automation of tasks.
- We have subfields focusing on learning, knowledge representation and reasoning, planning etc.



Task Automation: Vacuum Cleaner World



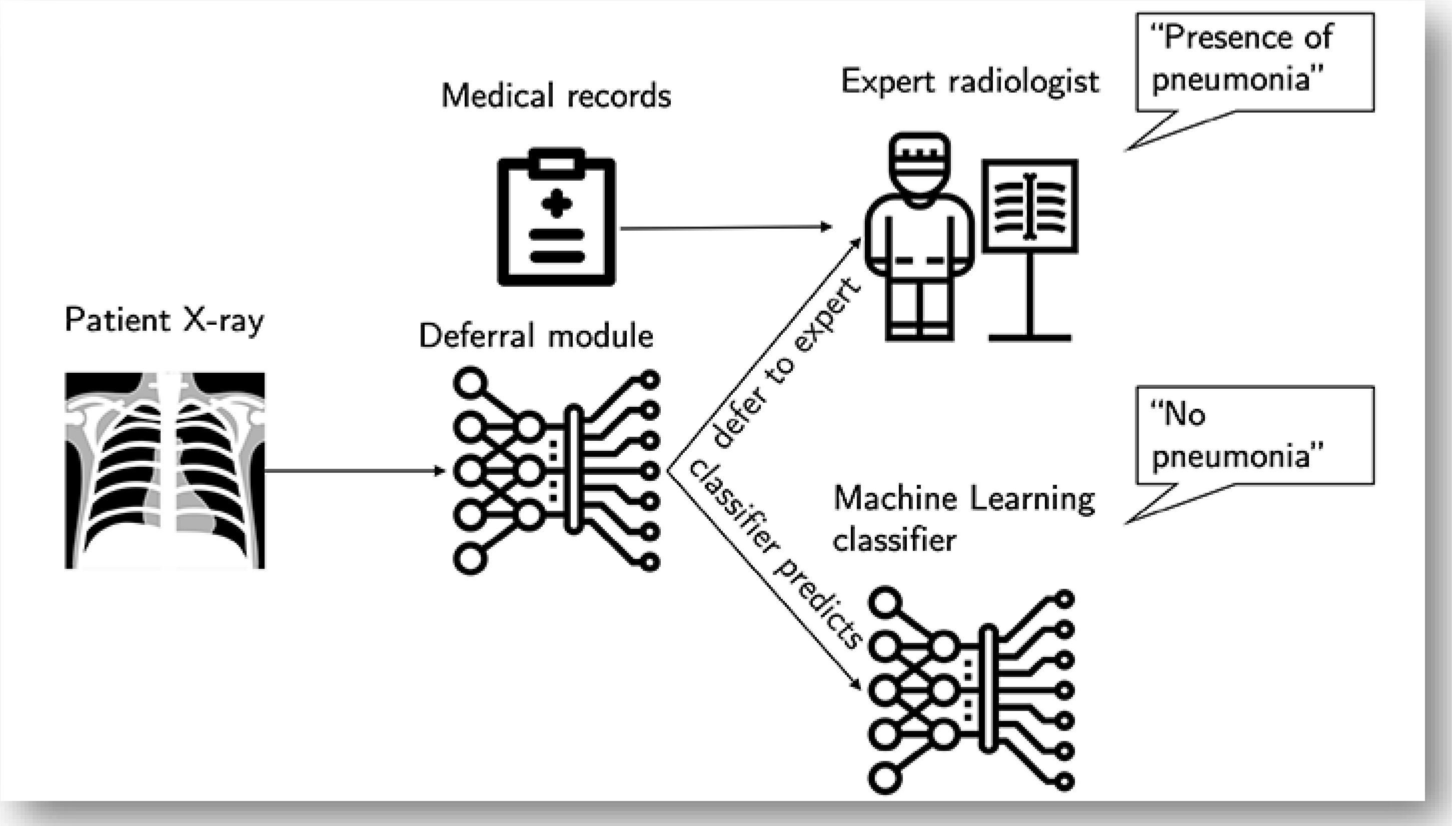
Example: Single state problems

- Let the world be consist of only 2 locations - Left and Right Box
- Intelligent agent → robot vacuum cleaner
- Sensors→ tell which sate it is in
- Known what each actions does
- Possible actions: *move left, move right, and suck.*
- **Goal**: we want all the dirt cleaned up.

  the goal is the state set $\{7, 8\}$.
- If the initial state is 5. Can calculate the action sequence to get to a goal state.

  [Right, Suck]

Example: A Human-in-the-Loop Approach

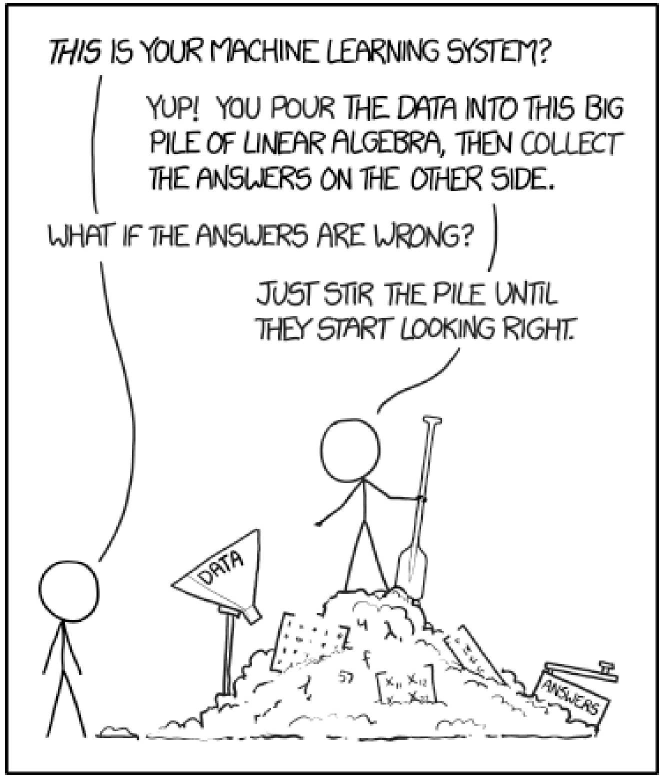


AI is everywhere!

- … from day-to-day tools to complex systems.
- Many domains involved:
  - Transport, marketing, healthcare, finance, insurance, security, science, education, agriculture, military, legal ….

(big) Data?

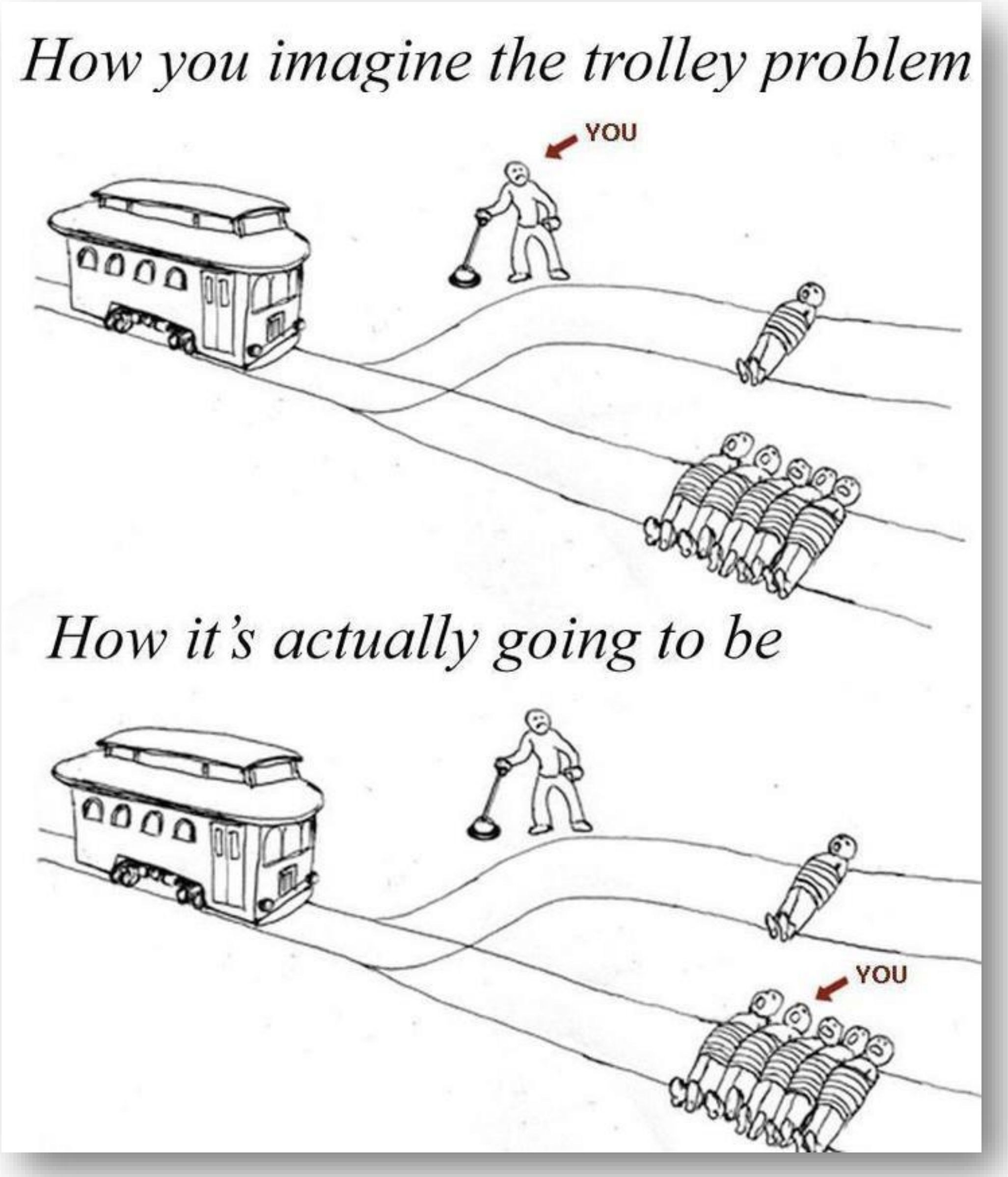
- Data IN, knowledge OUT
- We have enough computation power to:
  - Predict decisions
  - Model user behavior
  - …
- We should think of benefits and harms that an AI system could bring.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838/

Ethics?
Technology?

- Ethics focuses on good life.
  - A life with love, friendship, courage etc.
- It is best discussed as part of Philosophy.
  - Theoretical
  - Practical
- Technologies we develop have a big impact on power, justice and responsibility.


How you imagine the trolley problem
YOU
How it's actually going to be
YOU

What ethically significant harms and benefits can data present?

** based on Introduction to Data Ethics module (Part 1) developed by Shannon Vallor, Ph.D.*

What makes a harm or benefit ethically significant?

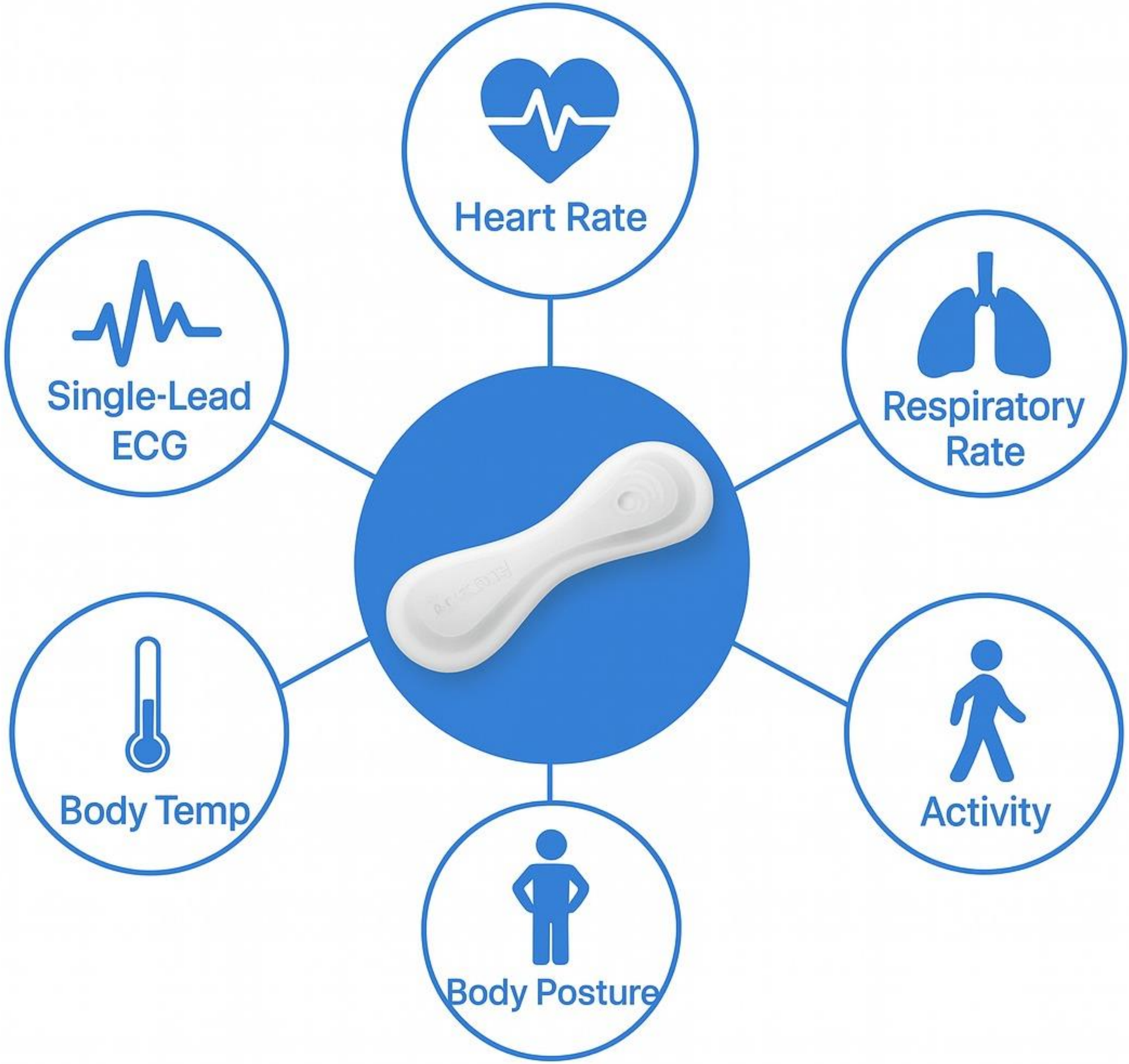
- We aim to have a 'good life'.
  - Not just ourselves, but as a society
- Ethically significant harm/benefit happens: "when it has a substantial possibility of making a difference to certain individuals' chances of having a good life, or the chances of a group to live well."
- Ethics implies 'human 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!

Ethically Significant Benefits of Data Practices

Human Understanding

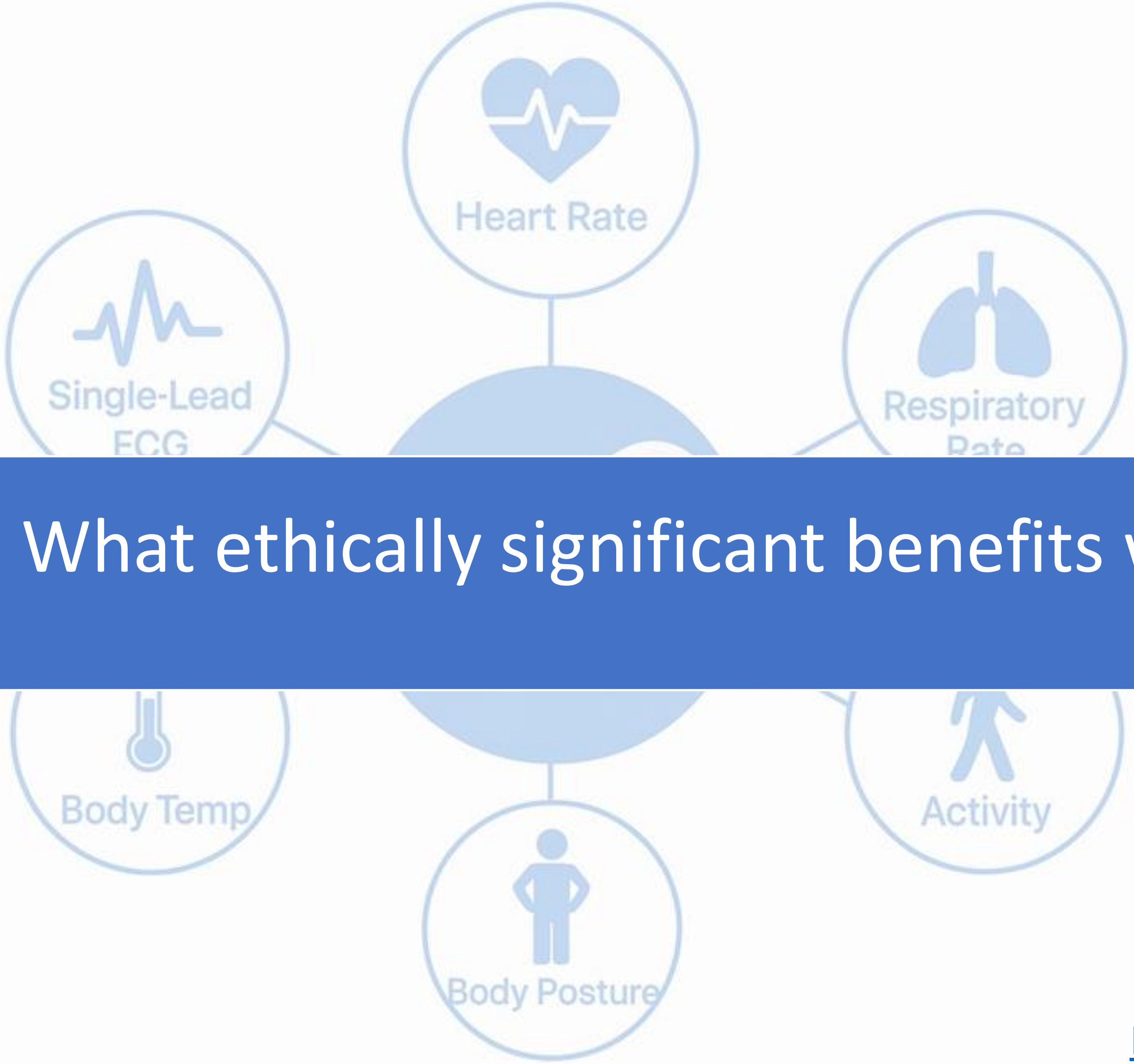
- We aim to understand the world, how it works to build better technology of the future.
- Complex systems vs smaller systems; we want to understand.
- We can identify (unseen) harms/needs/risks.
    - If we know a minority/marginalized group is being harmed, we may work for the benefit to a wider community.

Vital Patch: Remote Patient Monitoring



https://www.medibiosense.com/vitalpatch/

Vital Patch: Remote Patient Monitoring



What ethically significant benefits we could think of?

https://www.medibiosense.com/vitalpatch/

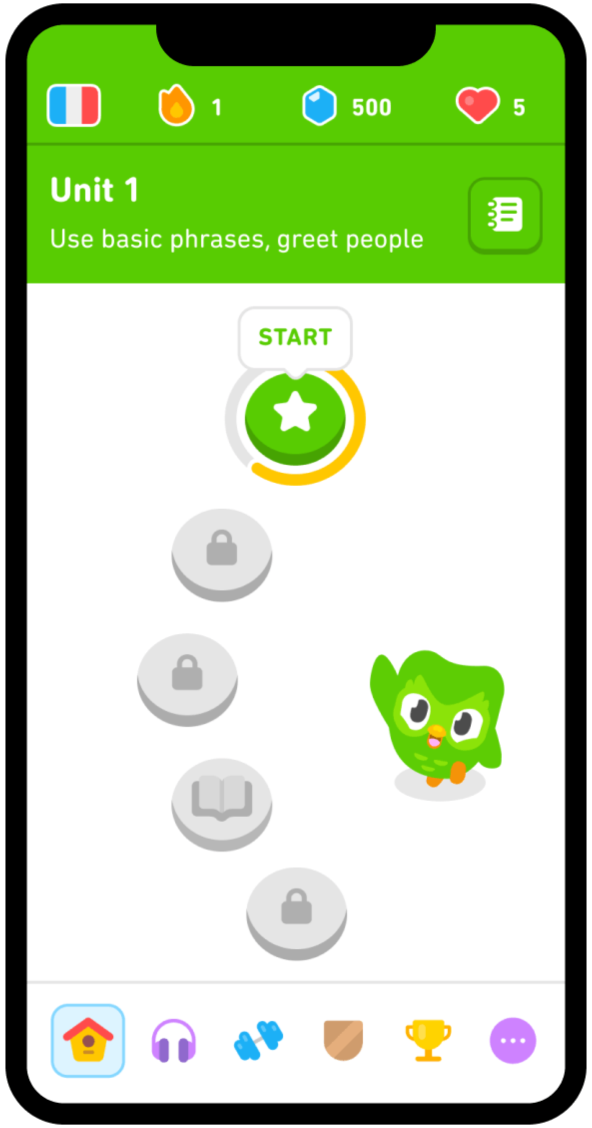
Predictive Accuracy and Personalization

- We can achieve good outcomes for specific individuals, groups.
- We can provide (real-time) feedback to user inputs.
- Context is key -> specific needs and circumstances
- Domains: search, ads (cookies), recommender systems, …

Duolingo

- An AI-based language learning platform.
- They use deep learning algorithms to personalize content.
- It gamifies the learning experience, hence more engagement from the users.

https://www.duolingo.com/

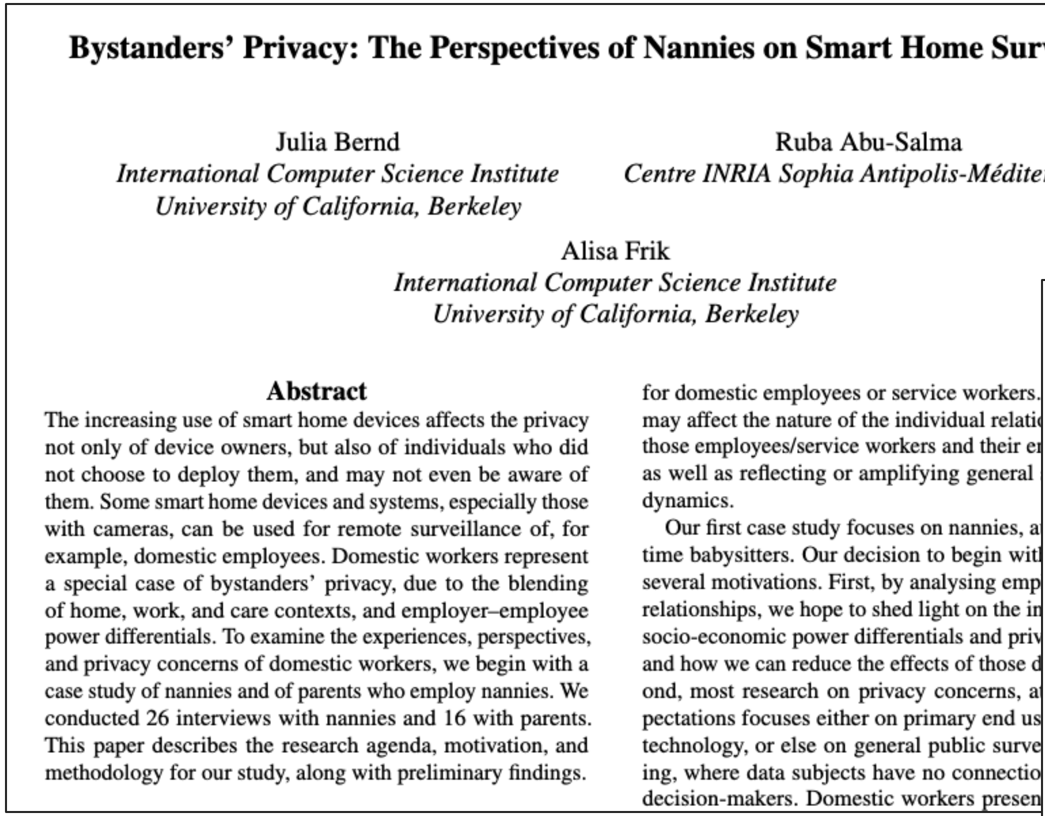


Ethically Significant Harms of Data Practices

Harms to Privacy & Security

- Everyone generates data.. We share data about ourselves as well as others.
- Anonymized datasets can be de-anonymized once merged with other datasets.
- Weakly enforced sets of data regulations and policies protect us from the reputational, economic and emotional harms.
- Harms can even be fatal (e.g., oppressive regimes).
- Data is also collected by deployed systems (e.g., IoT).

Privacy in Internet-of-Things Era



Bystanders' Privacy: The Perspectives of Nannies on Smart Home Surveillance

Julia Bernd
International Computer Science Institute
University of California, Berkeley

Ruba Abu-Salma
Centre INRIA Sophia Antipolis-Méditerranée

Alisa Frik
International Computer Science Institute
University of California, Berkeley

Abstract

The increasing use of smart home devices affects the privacy not only of device owners, but also of individuals who did not choose to deploy them, and may not even be aware of them. Some smart home devices and systems, especially those with cameras, can be used for remote surveillance of, for example, domestic employees. Domestic workers represent a special case of bystanders' privacy, due to the blending of home, work, and care contexts, and employer–employee power differentials. To examine the experiences, perspectives, and privacy concerns of domestic workers, we begin with a case study of nannies and of parents who employ nannies. We conducted 26 interviews with nannies and 16 with parents. This paper describes the research agenda, motivation, and methodology for our study, along with preliminary findings.

for domestic employees or service workers.
may affect the nature of the individual relati
those employees/service workers and their e
as well as reflecting or amplifying general
dynamics.

Our first case study focuses on nannies, a
time babysitters. Our decision to begin wit
several motivations. First, by analysing emp
relationships, we hope to shed light on the i
socio-economic power differentials and priv
and how we can reduce the effects of those d
ond, most research on privacy concerns, a
pectations focuses either on primary end us
technology, or else on general public surve
ing, where data subjects have no connectio
decision-makers. Domestic workers presen

Owning and Sharing: Privacy Perceptions of Smart Speaker Users

NICOLE MENG, University of Edinburgh, UK
DILARA KEKÜLLÜOĞLU, University of Edinburgh, UK
KAMI VANIEA, University of Edinburgh, UK

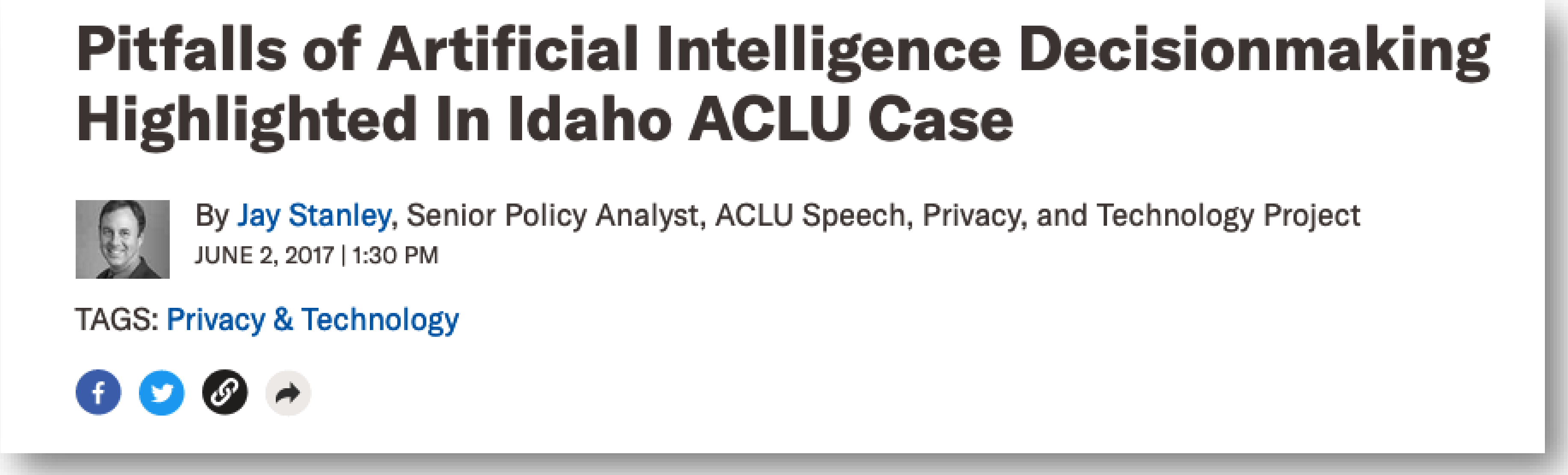
Intelligent personal assistants (IPA), such as Amazon Alexa and Google Assistant, are becoming increasingly present in multi-user households leading to questions about privacy and consent, particularly for those who do not directly own the device they interact with. When these devices are placed in shared spaces, every visitor and cohabitant becomes an indirect user, potentially leading to discomfort, misuse of services, or unintentional sharing of personal data. To better understand how owners and visitors perceive IPAs, we interviewed 10 in-house users (account owners and cohabitants) and 9 visitors from a student and young professionals sample who have interacted with such devices on various occasions. We find that cohabitants in shared households with regular IPA interactions see themselves as owners of the device, although not having the same controls as the account owner. Further, we determine the existence of a smart speaker etiquette which doubles as trust-based boundary management. Both in-house users and visitors demonstrate similar attitudes and concerns around data use, constant monitoring by the device, and the lack of transparency around device operations. We discuss interviewees' system understanding, concerns, and protection strategies and make recommendation to avoid tensions around shared devices.

Harms to Fairness and Justice (i)

- We want to be judged and treated fairly
- Biases that rest on falsehoods, sampling errors, and unjustifiable discriminatory practices are very common
- Implicit data biases are the most difficult ones to spot!
- Proxies can still be indicators of protected features such as race, gender etc.
  - zip code -> indicator of race or income

Harms to Fairness and Justice (ii)

- The harms can also be driven by:
  - Poor quality, mislabeled, error-riddled data
  - Inadequate design and testing of data analytics
  - Lack of training/auditing
- Some groups are affected by such data practices, and they lose their chance to live a 'good life'



Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case

By Jay Stanley, Senior Policy Analyst, ACLU Speech, Privacy, and Technology Project
JUNE 2, 2017 | 1:30 PM

TAGS: Privacy & Technology

- Disadvantaged group: 4000 Idahoans with developmental and intellectual disabilities
- The amount of assistance that they were being given by Medicaid program was being suddenly cut by 20 or 30 percent. An investigation takes place.
- It turns out that a magic Excel formula computes scores based on responses collected during an assessment review.
- They spend $50000 to test the system to understand the workings of the system.
- Many flaws detected: bad quality of data, partial use of historical data, incorrect statistics....

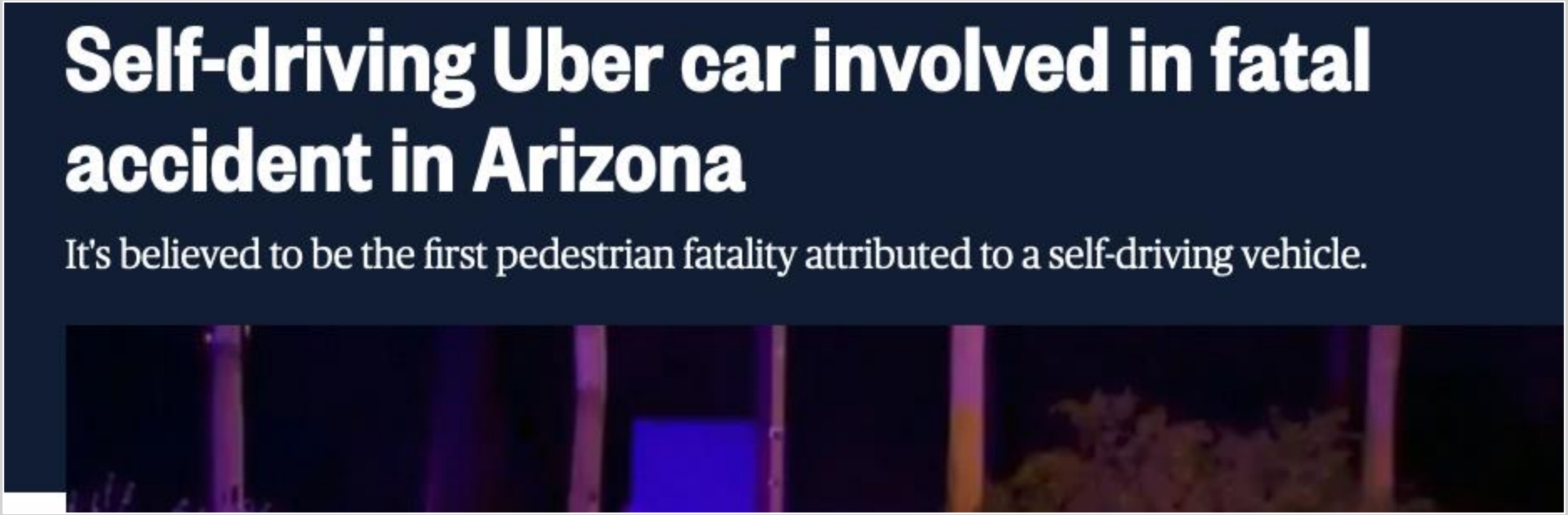
https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case

Harms to Transparency and Autonomy

- What is transparency?
  - It is the ability to see how a given social system or institution works
  - The focus in on understanding the outcome: "Why?"

- What is autonomy?
  - It is the ability to govern or steer the course of one's life
  - The focus is on control
  - Chances for a good life depend on you!

Who is responsible for this accident?



Self-driving Uber car involved in fatal accident in Arizona

It's believed to be the first pedestrian fatality attributed to a self-driving vehicle.

https://www.nbcnews.com/tech/innovation/self-driving-uber-car-involved-fatal-accident-arizona-n857941 (March 2018)

Case Study

- Now we are ready to think about Case Study 1 (page 13) in the following book:

An Introduction to Data Ethics by Shannon Vallor:

https://www.scu.edu/media/ethics-center/technology-ethics/IntroToDataEthics.pdf