



Case Studies in AI Ethics (CSAI)

2024/25

Meet your lecturer



Nadin KOKCIYAN

I am currently a Lecturer in Artificial Intelligence at the [School of Informatics, University of Edinburgh](#), and a Senior Research Affiliate at the [Centre for Technomoral Futures, Edinburgh Futures Institute](#).

I am:

- the director of the Human-Centered AI Lab ([CHAI Lab](#))
- a member of [Artificial Intelligence and its Applications Institute \(AIAI\)](#).
- a member of [Security and Privacy](#) group.
- affiliated with [Technology Usability Lab In Privacy and Security \(TULiPS\)](#).

Email: nadin.kokciyan at ed.ac.uk

Office: IF-2.10, School of Informatics, University of Edinburgh

Phone: +44 (0) 131 650 9993

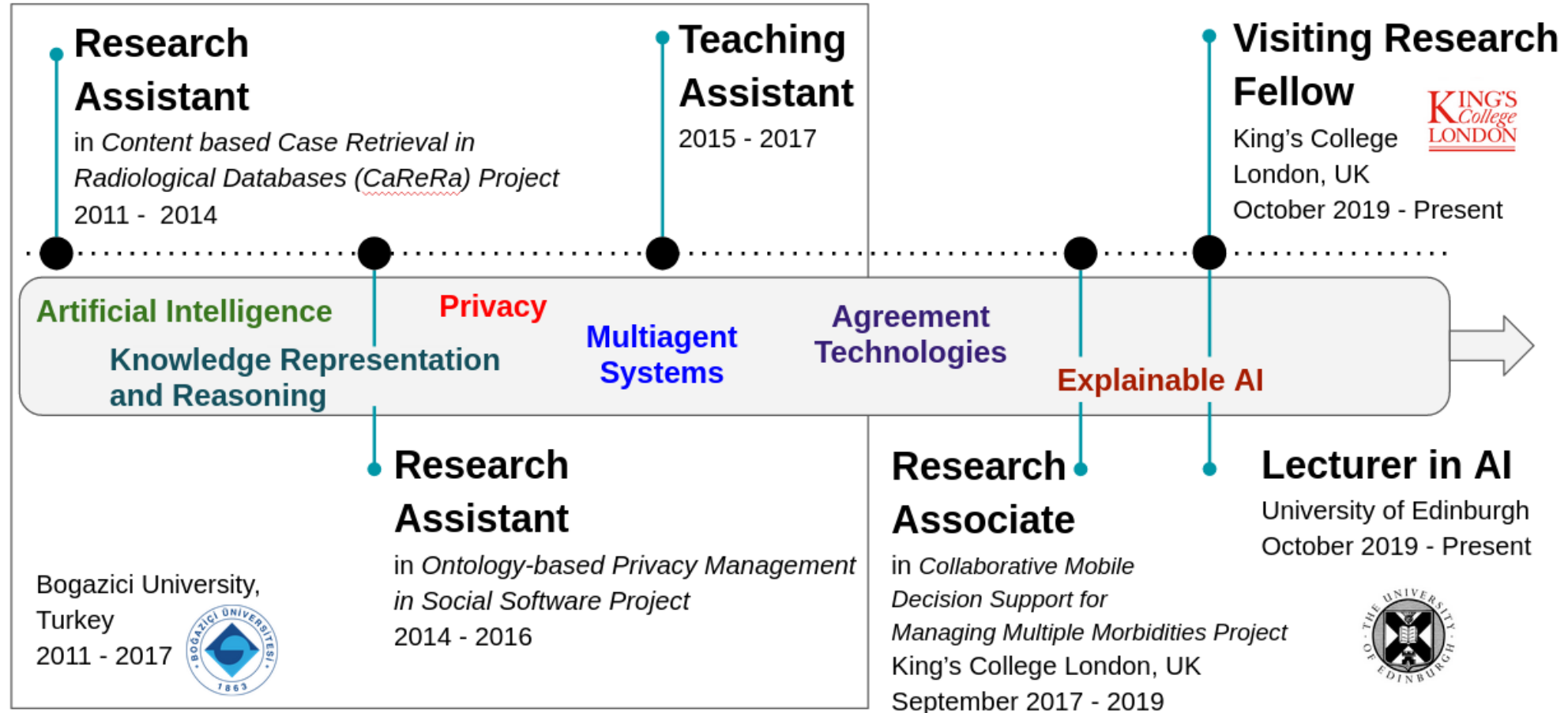
Research Interests

- Multiagent Systems
- Agreement Technologies (Argumentation and Negotiation)
- Privacy in Social Software
- AI Ethics, Explainable AI, Responsible AI

[Publications](#)

[CHAI Lab](#)

My Academic Journey

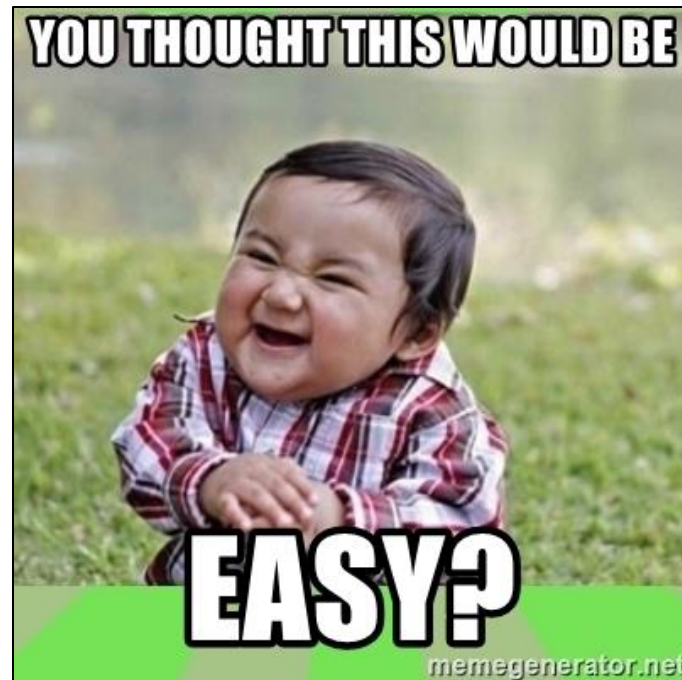


CSAI Teaching Support Team



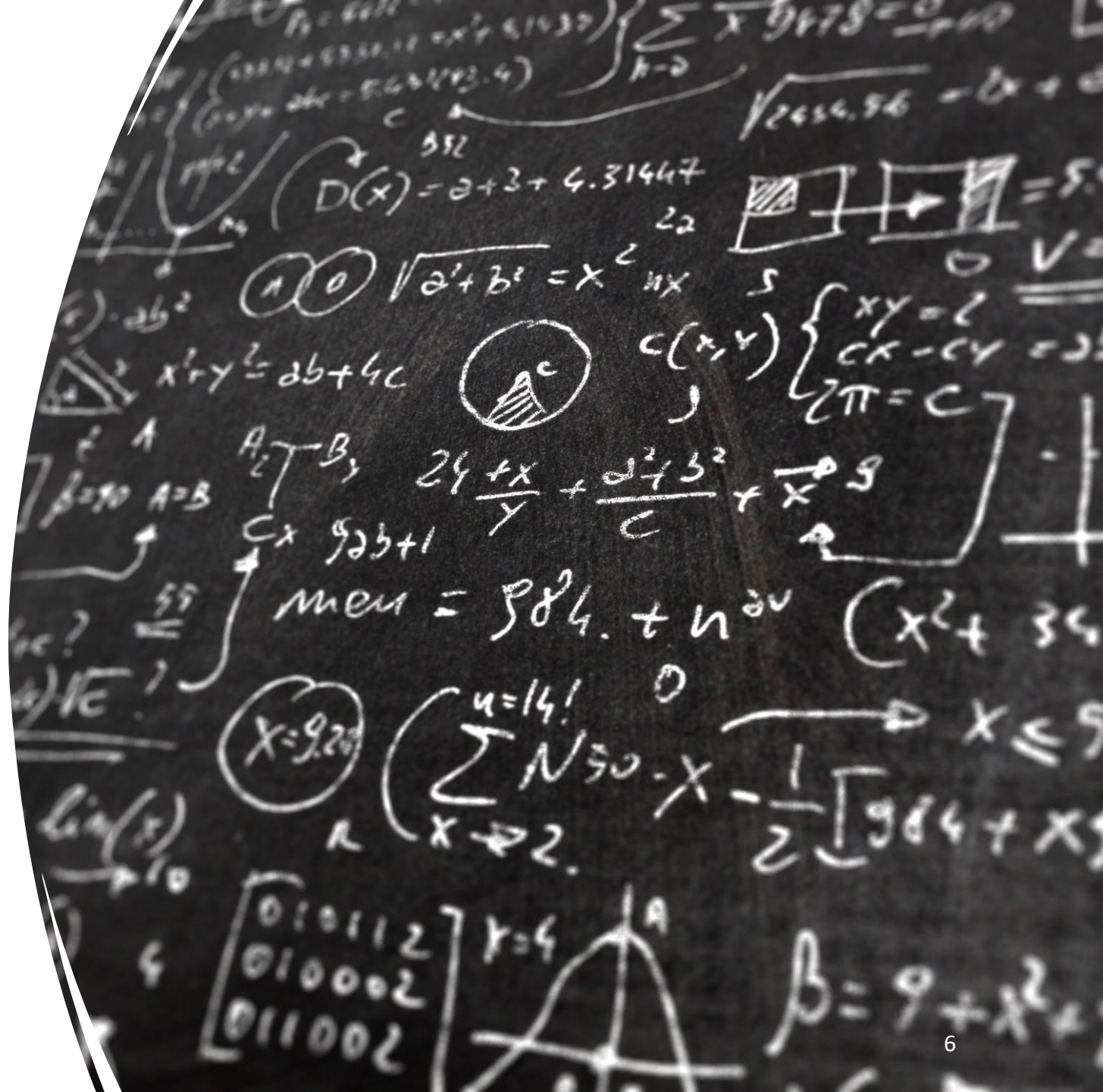
- Sydelle de Souza (Teaching Assistant)
- Fiona Smith (Tutor)
- Passara Chanchotisien (Marker)
- Sunnie Li (Marker)

First I want to convince you why we need this course...



AI?

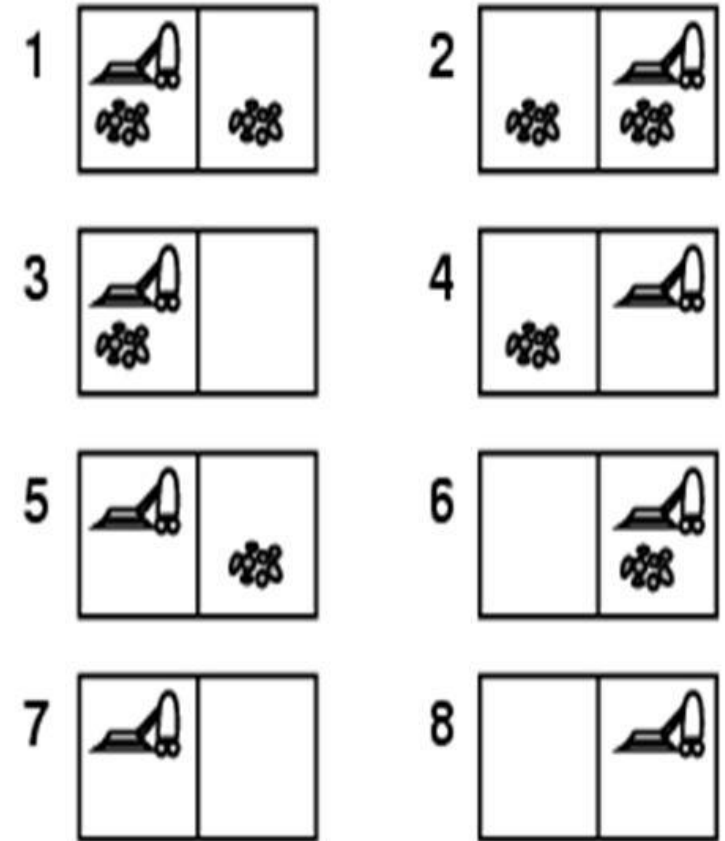
- Bellman (1978) defines AI as "the automation of activities that we associate with human thinking (i.e., cognitive activities)".
- Hence, the focus is on automation of tasks.
- We have subfields focusing on learning, knowledge representation and reasoning, planning etc.



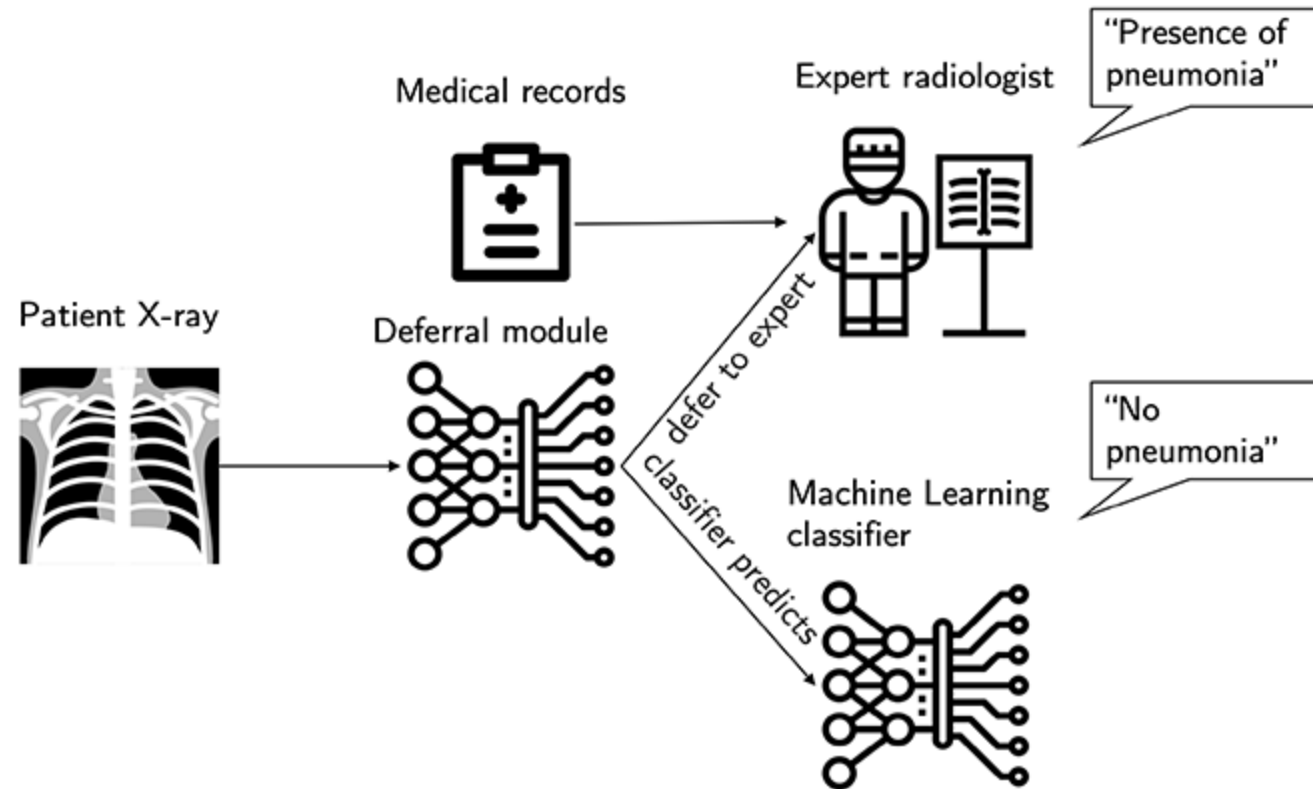
Task Automation: Vacuum Cleaner World

Example: Single state problems

- Let the world be consist of only 2 locations - Left and Right Box
- Intelligent agent \rightarrow robot vacuum cleaner
- Sensors \rightarrow 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: An Automated Diagnostic Tool



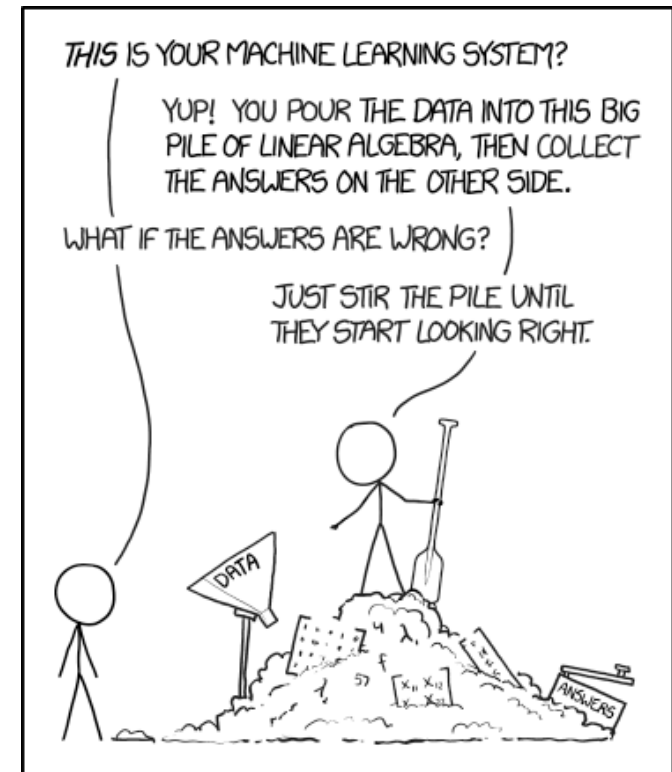
AI is everywhere!

- ... from day-to-day tools to complex systems.
- Many domains involved:
 - Transport, marketing, healthcare, finance, insurance, security, science, education, agriculture, military, legal
- Big tech firms know very well how to be part of our lives!



(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.



<https://xkcd.com/1838/>

ARTICLE

Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists

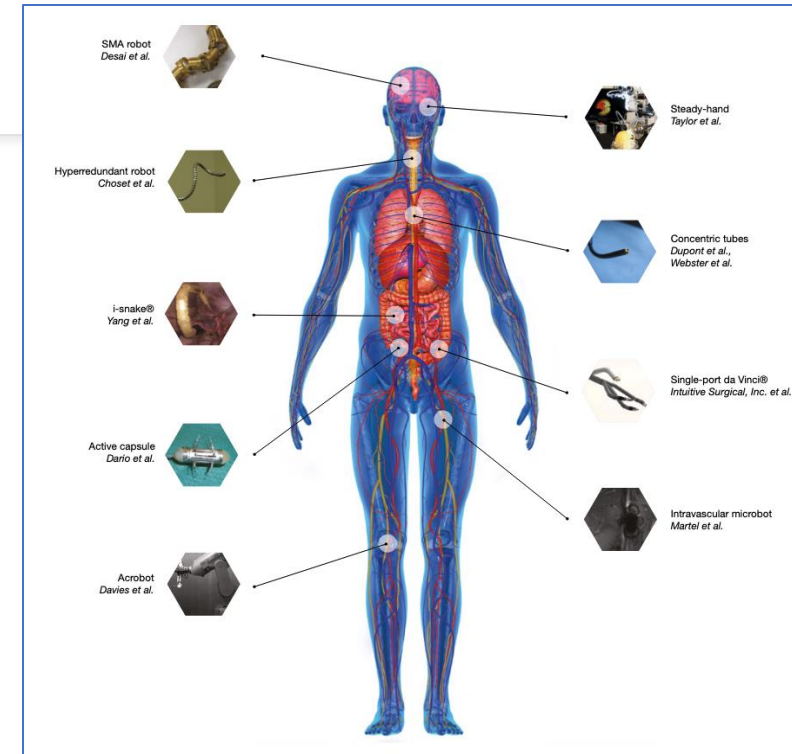
Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H. Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, Matthew G. Wallis, Ingvar Andersson, Sophia Zackrisson, Ritse M. Mann, Ioannis Sechopoulos

Features

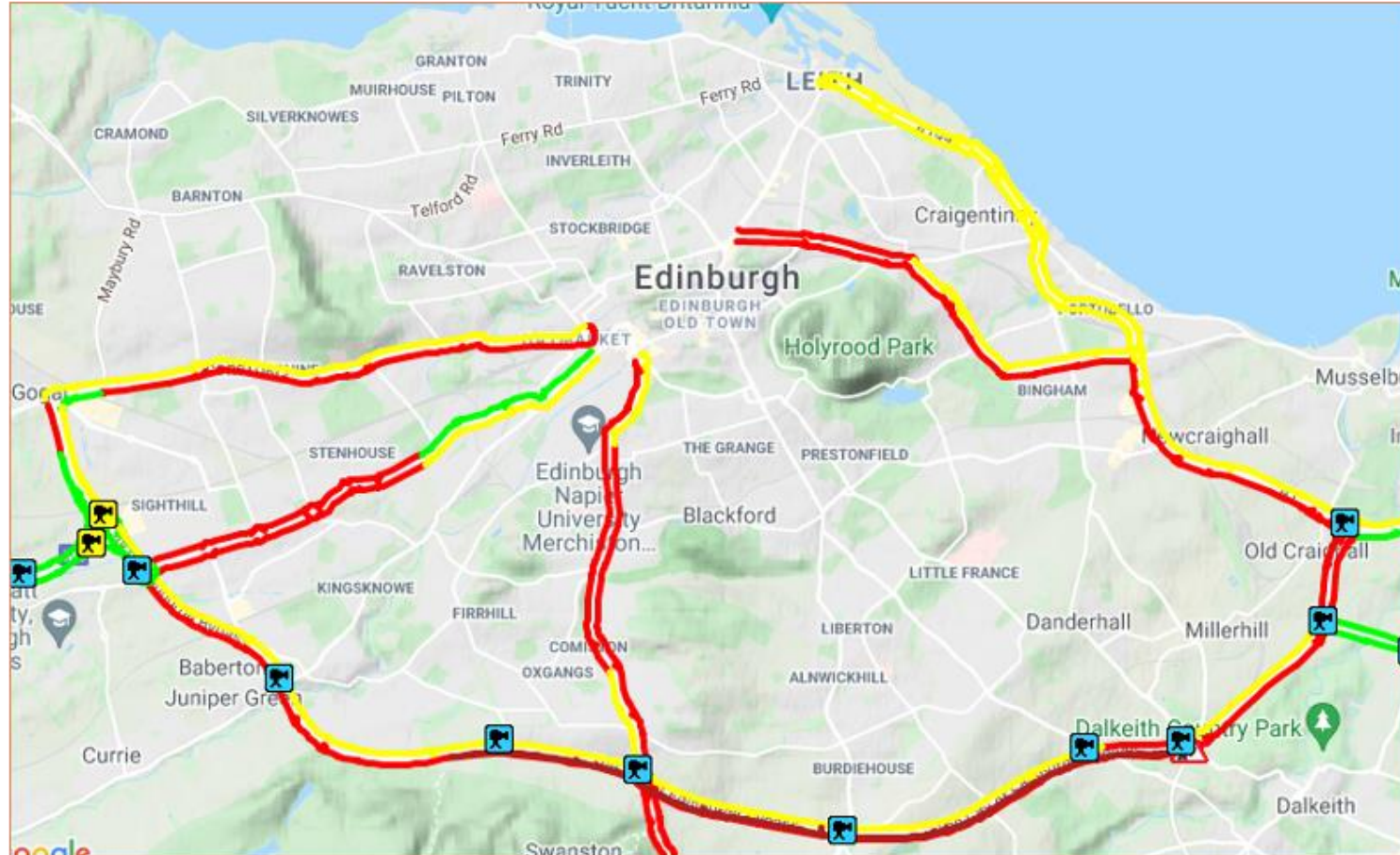
VitalPatch monitors a total of eight vital signs:



The Vital Patch is a health monitoring device in the growing field of Tele-Health. Never before has such a small, elegant device provided so much valuable information for physicians and nurses. This state-of-the-art biosensor monitors eight physiological measurements continuously, in real time. Clinical-grade accuracy without the hassle of traditional monitoring equipment. The best things do come in small packages.



Happy people, happy environment...



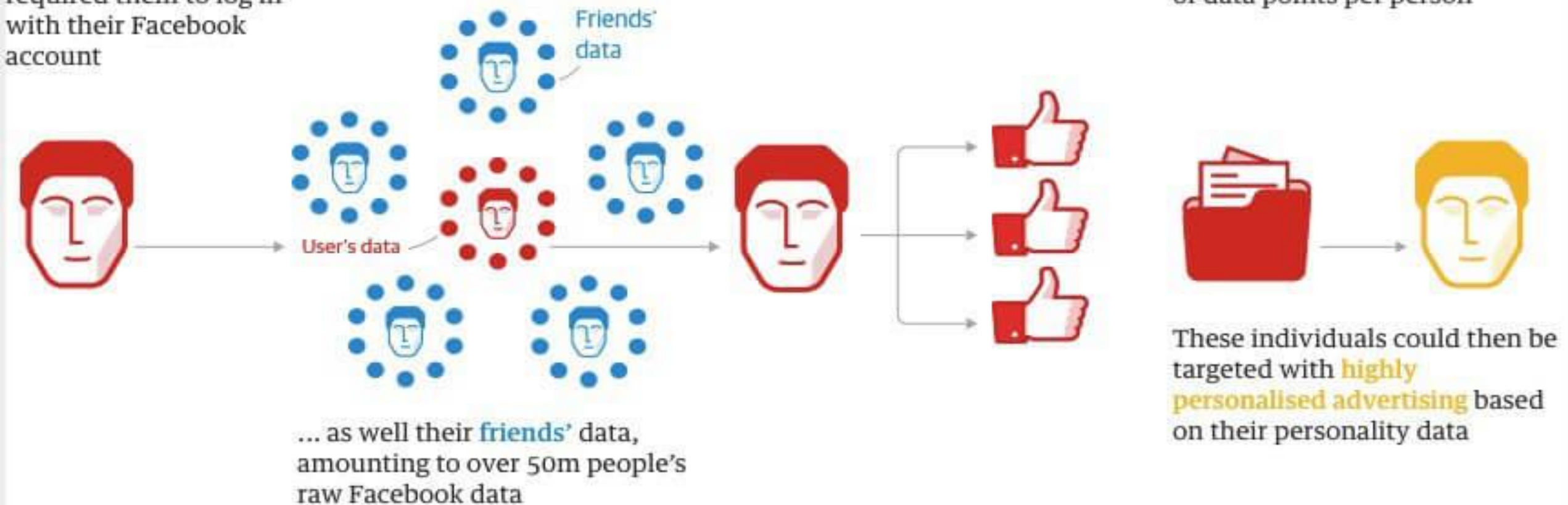
Cambridge Analytica: how 50m Facebook records were hijacked

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



#disarmICE

In 2017, Palantir software allowed ICE to launch an operation that targeted and arrested family members of children who crossed the border, leading to 443 arrests.

Ethical Issues: deporting migrants, refugees, and asylum seekers, separating families, keeping children in detention...

"The question isn't whether you're undocumented — but rather whether a flawed algorithm thinks you look like someone who's undocumented."

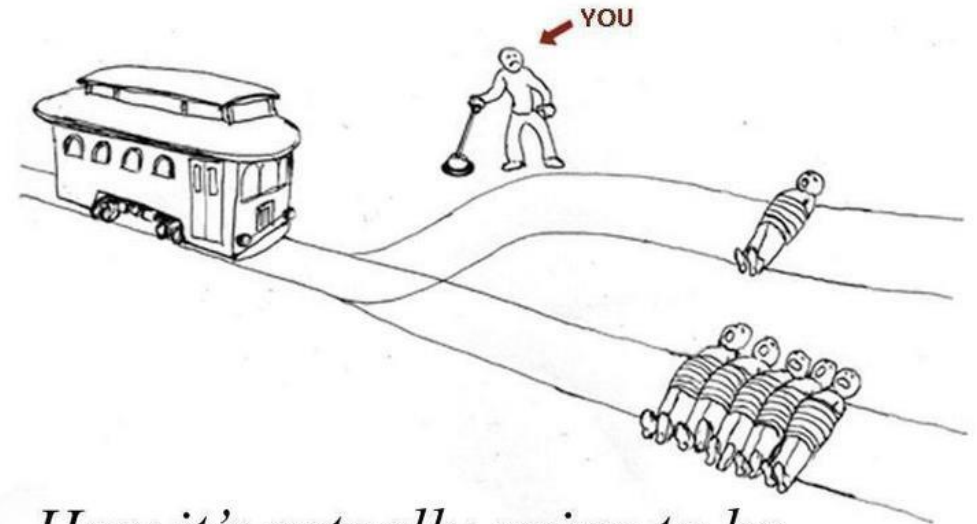
Alvaro Bedoya,
the founding director of Georgetown Law's Center on Privacy & Technology.

Ethics?

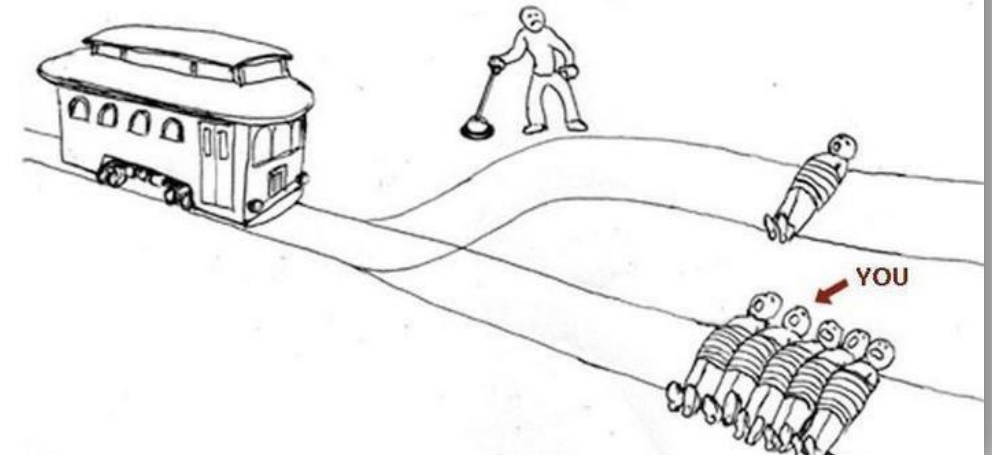
Technology?

- 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

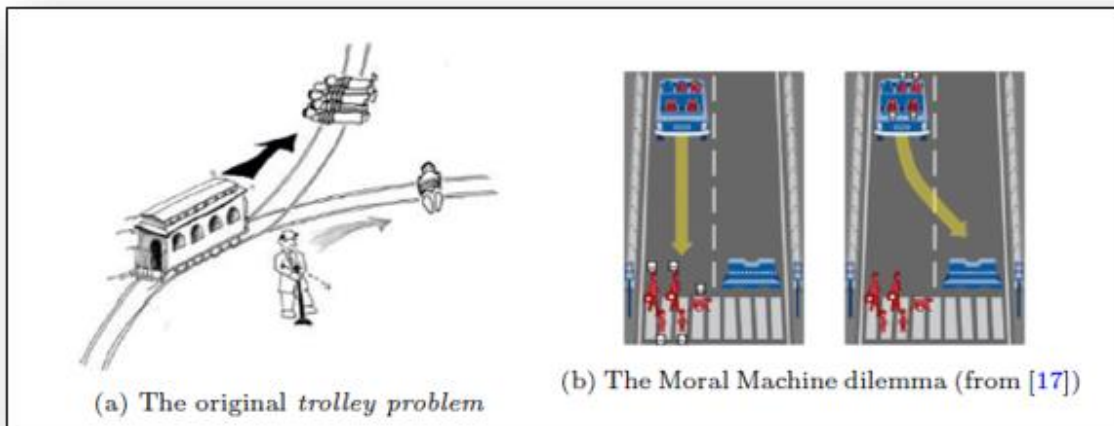


How it's actually going to be



We have a new trolley problem!

- Should self-driving cars have built-in ethics constraints? What constraints? How to identify these?



A Voting-Based System for Ethical Decision Making

Ritesh Noothigattu¹, Snehal Kumar 'Neil' S. Gaikwad², Edmond Awad², Sohan Dsouza²,
Iyad Rahwan², Pradeep Ravikumar¹, and Ariel D. Procaccia¹

¹School of Computer Science, Carnegie Mellon University

²The Media Lab, Massachusetts Institute of Technology

Abstract

We present a general approach to automating ethical decisions, drawing on machine learning and computational social choice. In a nutshell, we propose to *learn* a model of societal prefer-

Some other ethical concerns...

- How much of our decisions we want to delegate to AI?
- The COMPAS algorithm is **highly controversial**, the algorithm's false positives are disproportionately black (Fry 2018).
- Predictive policing: where crimes are likely to occur and who might commit them
 - Specific socioeconomic or racial groups may be targeted by **police surveillance**.

Some other ethical concerns...

HASTA LA VISTA, BABY —

Microsoft terminates its Tay AI chatbot after she turns into a Nazi

Setting her neural net processor to read-write was a terrible mistake.

ARS STAFF - MAR 24, 2016 2:28 PM UTC



How deepfakes are impacting society

Fake online video and audio content has become a powerful tool for spreading political misinformation and harming personal reputations.

By Morgan Currie, Lecturer in Data & Society at the School of Social and Political Science

CSAI: Course Introduction



Learning Outcomes

- Understand data ethics and arising issues (e.g., bias, fairness, privacy) in AI systems.
- Explain and provide examples of how AI systems can play a critical role in decision making.
- Analyse case studies to identify and mitigate potential risks considering legal, social, ethical or professional issues.
- Apply ethical methodologies in the design of responsible AI systems.

CSAI: Course Content

Data Ethics, Machine Ethics, AI Ethics

Data Ethics

- What are ethically significant harms and benefits?
- Common ethical challenges data practitioners and users face.
 - Data collection
 - Data storage, security
 - Data hygiene
 - Identifying/addressing bias
 - ...

** based on Introduction to Data Ethics module (several chapters) by Prof Shannon Vallor*

Machine Ethics

- How to automate **moral reasoning** for computational agents?
- Four different types of agents*:
 - Ethical-impact agents
 - Implicit ethical agents
 - **Explicit ethical agents**
 - Full ethical agents

* Moor, James H.: *The Nature, Importance, and Difficulty of Machine Ethics*. In: *IEEE Intelligent Systems* 21 (2006), Juli, Nr. 4, S. 18–21.

Spec in YAML

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

1  description: The Rescue Robot Dilemma
2  actions: [a_save_h1, a_save_h2, a_remain_inactive]
3  background: [b_save_people]
4  consequences: [saved_h1, discomfort_h1, saved_h2, discomfort_h2]
5  mechanisms:
6      saved_h1: And("b_save_people", "a_save_h1")
7      discomfort_h1: a_save_h1
8      saved_h2: And("b_save_people", "a_save_h2")
9      discomfort_h2: a_save_h2
10 utilities:
11     saved_h1: 10
12     discomfort_h1: -4
13     saved_h2: 10
14     discomfort_h2: -4
15     Not('saved_h1'): -10
16     Not('discomfort_h1'): 4
17     Not('saved_h2'): -10
18     Not('discomfort_h2'): 4
19 intentions:
20     a_save_h1: [a_save_h1, saved_h1]
21     a_save_h2: [a_save_h2, saved_h2]
22     a_remain_inactive: [a_remain_inactive]
23
```


AI Ethics

We build (semi/fully) automated systems that interact with people.

These systems are heterogeneous and consist of various components.

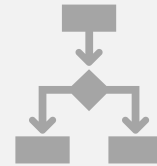
How to ensure that these systems overall do not harm people?

AI Ethics



Fairness,
Accountability, and
Transparency (FAccT)

How to put
human in
control of AI
systems?



Explainable AI (XAI)

Could AI be
more
explainable?



Responsible AI

How to
regulate and
deploy AI?

Course Structure

Lectures, Tutorials, Courseworks, Exam

Lectures

- Each lecture:
 - Happens in person (fingers-crossed!)
 - Covers content on the topic of the week
 - Includes class discussion
- We will have two case-study weeks
- You will need to:
 - Do weekly readings/watch videos/experiment with web sites

Active participation is required




Tutorials

- We will have two tutorials.
- The first one is on **Week 4**, the second one is on **Week 7**.
 - First tutorial will focus on **critical thinking** and **active discussion**. We will work on a case study as a group.
 - Second tutorial will focus on a **practical example**.

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** (1500-2000 words) based on the outline submitted as CW1.

| Assignment | Released | Submit By | Feedback By | Extensions | Notes |
|--|------------|----------------------------|-------------|--|--|
| CW1 Essay Outline (Group) Worth: 0% Effort: 4 Hours | 27/01/2025 | Monday 10/02/2025 12:00 | 03/03/2025 | Rule 2: No Extensions and no ETAs permitted. Late individual submissions score zero. Penalties applied to late group work submissions. | Involves group work. via Gradescope |
| CW2 Essay (Individual) Worth: 40% Effort: 10 Hours | 24/02/2025 | Monday 17/03/2025 12:00 | 07/04/2025 | Rule 1: Extensions (4 days) and ETAs (7 days) permitted. Penalties applied to late submissions. | via Gradescope |

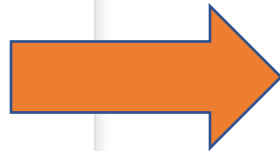


Exam (60%)

- Case Studies
- Application of ethical frameworks to specific cases
- Questions about data ethics, machine ethics and AI ethics

Course Structure: Questions

- We will use **Piazza** for active discussion.
- If you decide to send me an email, use the hashtag #CSAI in your subject line.
- This will be me for sure





Questions?

