Case Studies in AI Ethics (CSAI)

2023/24
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
My Academic Journey

Research Assistant
in Content based Case Retrieval in Radiological Databases (CaReRa) Project
2011 - 2014

Research Assistant
Artificial Intelligence
Knowledge Representation and Reasoning
Bogazici University, Turkey
2011 - 2017

Teaching Assistant
Privacy
Multiagent Systems
Agreement Technologies
2015 - 2017

Visiting Research Fellow
Visiting Research Fellow
King's College London, UK
October 2019 - Present

Research Associate
Explainable AI
in Collaborative Mobile Decision Support for Managing Multiple Morbidities Project
King's College London, UK
September 2017 - 2019

Lecturer in AI
2019 - Present

University of Edinburgh
CSAI Teaching Support Team

Eddie Ungless, Fiona Smith, Ana Deligny
First I want to convince you why we need this course...
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 state 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]

Solving Problems by Searching®
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


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

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

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.

*ICE: Immigration and Customs Enforcement
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**.
We have a new trolley problem!

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

https://www.moralmachine.net/
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.

**ABS 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*

[https://www.ed.ac.uk/impact/opinion/how-deepfakes-are-impacting-society](https://www.ed.ac.uk/impact/opinion/how-deepfakes-are-impacting-society)
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

Spec in 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]
Example in action: Jupyter Notebook

Importing the required modules

The evaluation will be done with respect to two moral principles: the principle of double effect and the utilitarian principle.

```python
In []: from ethicis.com.principles import DoubleEffectPrinciple, UtilitarianPrinciple, DoNoHarmPrinciple
      from ethicis.com.semantic import CausalModel

setting up the models

```python
In []: models = []
  m1 = CausalModel("/examples/rescue-robot.yaml", C_a_save_H1: 1, "a_save_H2": 0, "a_remain_inactive": 0, "b_save_pe
  m1._str_ = "save H1"
  m2 = CausalModel("/examples/rescue-robot.yaml", C_a_save_H1: 0, "a_save_H2": 1, "a_remain_inactive": 0, "b_save_pe
  m2._str_ = "save H2"
  m3 = CausalModel("/examples/rescue-robot.yaml", C_a_save_H1: 0, "a_save_H2": 0, "a_remain_inactive": 1, "b_save_pe
  m3._str_ = "remain inactive"

models.extend([m1,m2,m3])
```

defining models as alternatives of each other

```python
In []: for n in models:
    n.alternatives = models
```

evaluation of the models

```python
In []: for n in models:
    res1 = n.evaluate(DoubleEffectPrinciple)
    res2 = n.evaluate(UtilitarianPrinciple)
    res3 = n.evaluate(DoNoHarmPrinciple)
    print("Principal of Double Effect: ", res1, "\nUtilitarianism: ", res2, "\nDo-No-Harm: ", res3, "\n")
```

explanation of the models

```python
In []: print(m1.explain(DoubleEffectPrinciple))
In []: print(m1.explain(UtilitarianPrinciple))
In []: print(m1.explain(DoNoHarmPrinciple))
```
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. We will use AI Fairness 360 toolkit to analyse real world data.
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. The outline will not be more than two pages.

• **CW2: Essay (40%)** - This coursework is an individual assessment. Each student will write an essay (1500-2000 words) based on the outline submitted as CW1.
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?