THE UNIVERSITY *of* EDINBURGH

Case Studies in Responsible Natural Language Processing – CS-NLP

Prof Frauke Zeller, Ella Markham

Week 1 – Introduction to the Course

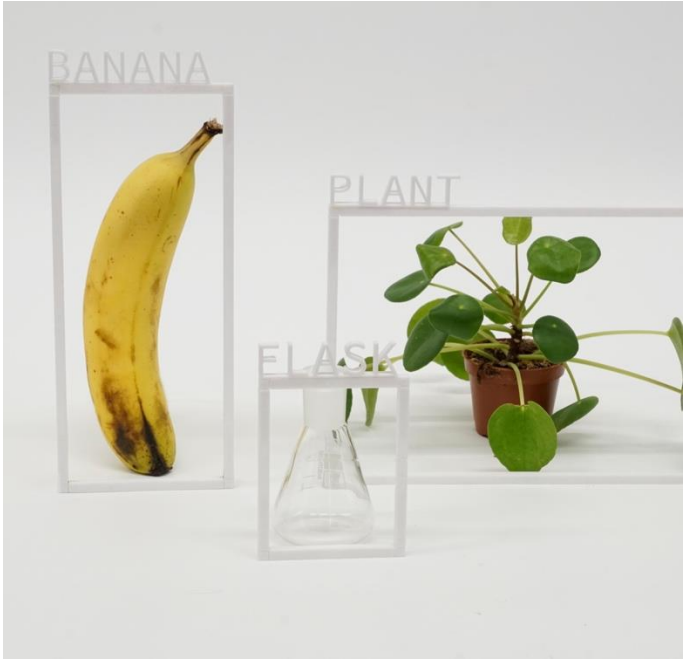
- Who are we
- What the course is about
- What do you expect, how can we work together
- Assessments
- AOB

This course will deliver key aspects of the responsible research and innovation (RRI) training provided by the CDT in Designing Responsible NLP.

It will build on courses on **legal, social, and ethical aspects of AI and NLP** that you took in year 1.

This course is **interdisciplinary** – and interdisciplinarity requires **exchange of ideas and communication**.

This means we will discuss our group/individual progress in class on a **weekly** basis to be able to better reflect on our own disciplinary bias and arrive at new/other ideas.



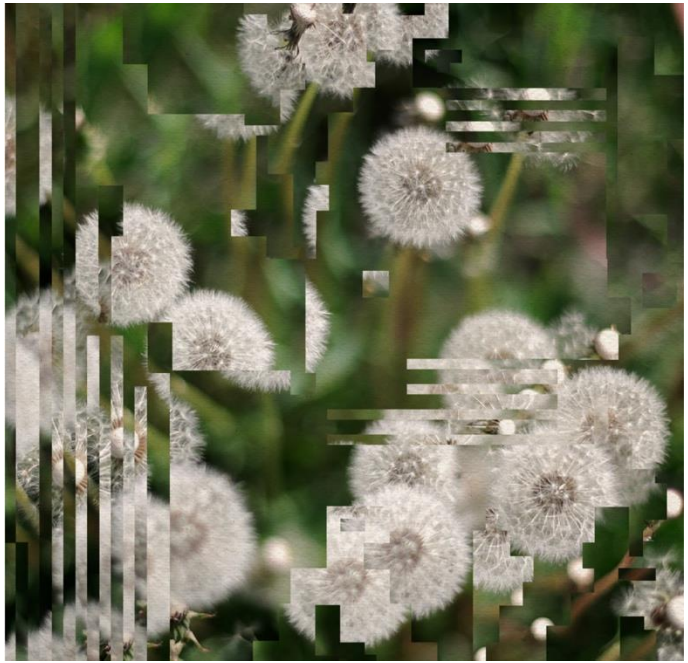
Max Gruber / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

The course will focus on applying responsible NLP principles in practice. This will take two forms:

(1) you will report and reflect on your year 1 learning in the area of legal, social, and ethical aspects of AI and NLP,

(b) you will work on case studies in form of real-world AI audits that draw out insights for the development and practical application of principles of responsible NLP.

Thus, the course will enable you to practice responsible research and innovation in action.



Lone Thomasky & Bits&Bäume / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Learning Outcomes:

On completion of this course, the student will be able to

1. critically evaluate the literature on legal, social, and ethical aspect of NLP

2. working with partners, analyse legal, social, and ethical implications of deploying NLP technology across various application domains

3. design potential solutions to legal, social and ethical problems, combining engineering and design thinking

Coursework:

There will be two pieces of coursework:

1. Seminar presentation on a topic in responsible NLP (35%)

2. Report on an analysis of a case study/AI Audit suggested by a CDT partner (65%)

Both assignments are group work.

Nadia Nadesan & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Assignment 1:

Three group presentations on the main themes you have taken courses on in year 1: legal, social and ethical aspects of AI and NLP.

You are expected to instruct/teach the other students on those thematic areas they don't have much background yet.

Assignment 1 – the Nitty Gritty:

- You will work in groups of 4-3 persons, 3 groups in total

- Each group with one focal point: Legal, Social or Ethical aspects of AI and NLP

- Presentation will be in class, 35-40 minutes plus 20 minutes guided discussion

- The presentation will also contain **practical group/individual work** (about 15 minutes top), which has to be integrated in the presentation (within the overall 35-40 minutes).

- Each group will submit the slides of the presentation on day of presentation.

- Each group will also submit a short report on how the work was split up among the group members, what worked, what did not work – 500 words max. Each group member will have to participate in the presentation but this can be of different lengths.

Assignment 1 – the Nitty Gritty:

The assignment will be a pass/fail activity. Formative feedback will be provided on:

- overall performance, time management
- pacing, voice projection, natural/free speaking
- composition and design of slides
- content: adequacy, 3 academic sources, etc.
- discussion moderation
- in-class work organisation and management
- overall group work management

Peers will also be expected to provide feedback.

Assignment 2:

Presentation and report on an analysis of a case study/Audit suggested by a CDT partner.

- You will work in groups of 3-4
- You will choose one of the offered cases to conduct an AI “audit”.

- The terms and details of the audit will be defined together with the partner and students and instructor.

Nadia Nadesan & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

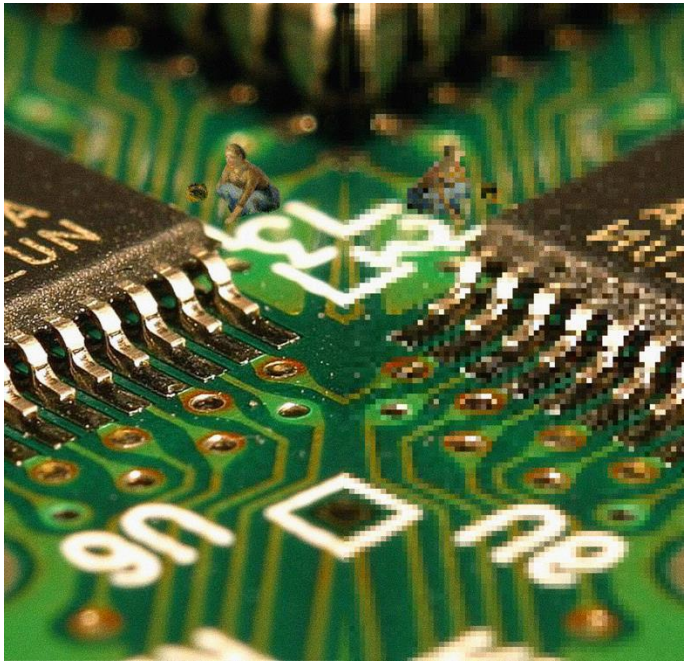
Assignment 2 – the Nitty Gritty:

- You will present in groups the results of the audit during the final class of the term (potentially with some of the CDT partners present).

- Presentation slides should be submitted via Learn a week before the presentation date.

- You will write up the results in your group and submit a written report (2000 words, including references). Additional material will be ignored by the markers.

WHAT IS AN AUDIT IN AI?



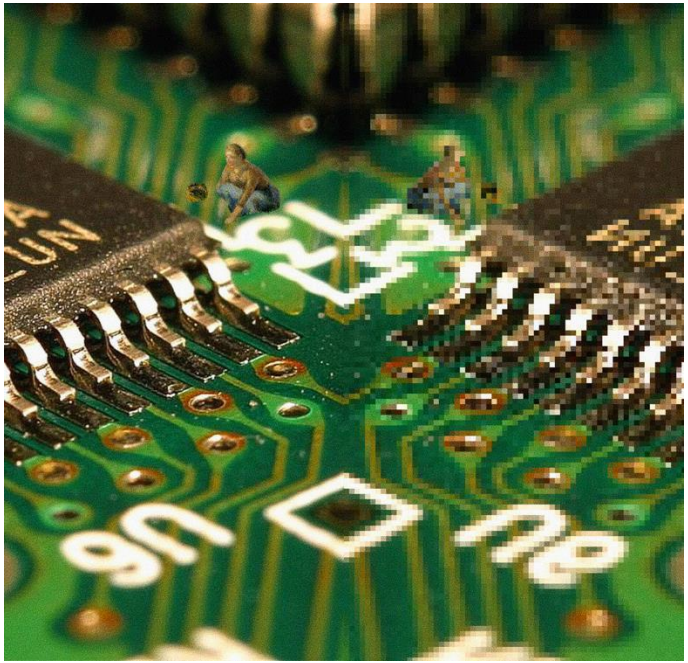
Elise Racine & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

WHAT IS AN AUDIT IN AI?

We can discuss 'AI Audit' as both a **research tool** and as an **object of research**:

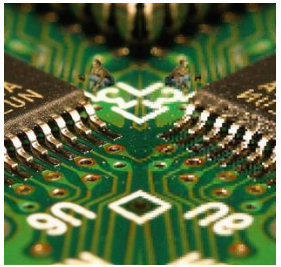
As a **tool** means that companies/organisations/individuals use AI in auditing processes.

As an **object of research** means that we look at AI systems in organisations/institutions/etc. and conduct an audit in terms of the technical, financial, ethical, etc. performance.



Elise Racine & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

WHAT IS AN AUDIT IN AI?



"As [AI] audits have proliferated, the meaning of the term has become ambiguous, making it hard to pin down what audits actually entail and what they aim to deliver" (Vecchione et al., 2021, p. 1; in Mökander, 2023, p.4)

"In the social sciences, the term 'audit study' refers to a research method, specifically a type of field experiment, which is used to examine individuals' behaviour or the dynamics of social processes" (Gaddis, 2018; in Mökander, 2023, p.7).

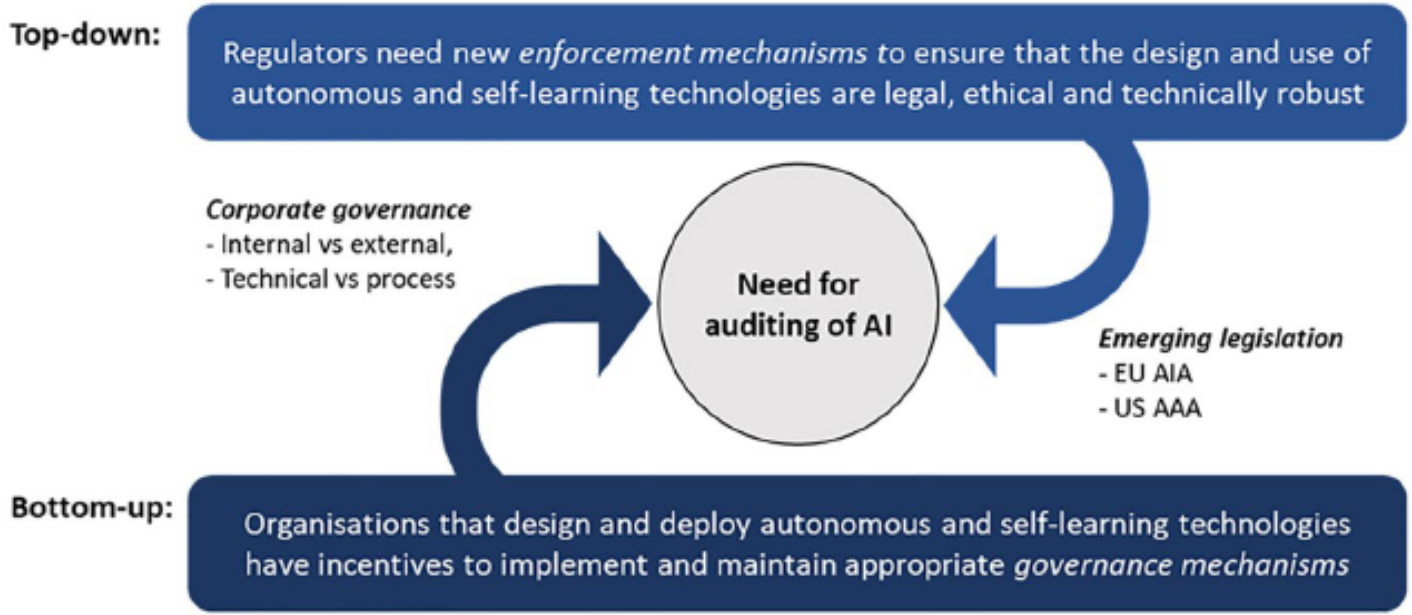
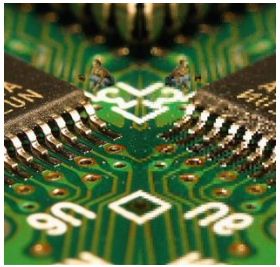


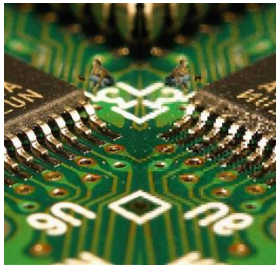
Fig. 2 The need to audit AI systems is underpinned by both top-down and bottom-up pressures

WHAT IS AI Governance?

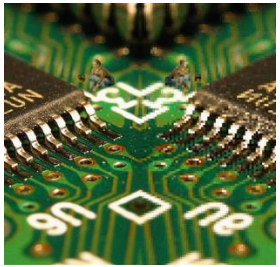


WHAT IS AI Governance?

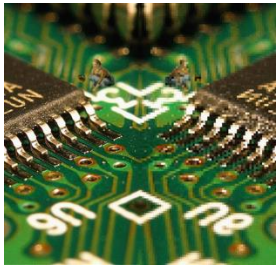
“AI governance is a system of rules, practices, processes, and technological tools that are employed to ensure that an organization’s use of AI systems aligns with the organization’s strategies, objectives, and values (M.ntym.ki et al., 2022, p.2).”



WHAT IS AN ALGORITHM AI Audit?



WHAT IS AN ALGORITHM AI Audit?



"[an algorithm audit is] a method of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs in order to draw inferences to its opaque inner workings (Metaxa et al., 2021, p.18)."

However, "It is not just about checking the algorithm itself and the management measures surrounding it, but also paying attention to the data used, the methods used in the development and the optimization of the algorithm. These aspects of management, process, and content should also be part of the assessment framework and thus the audit approach (Jager & Westhoek, 2023, p.145)."

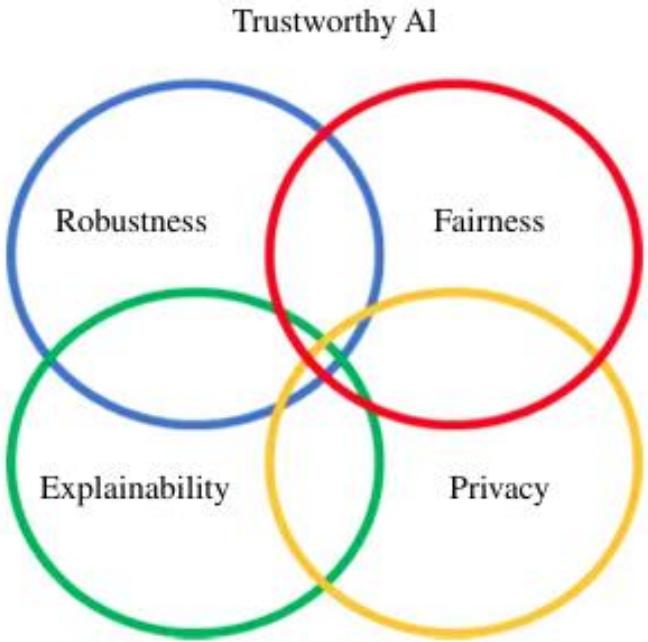
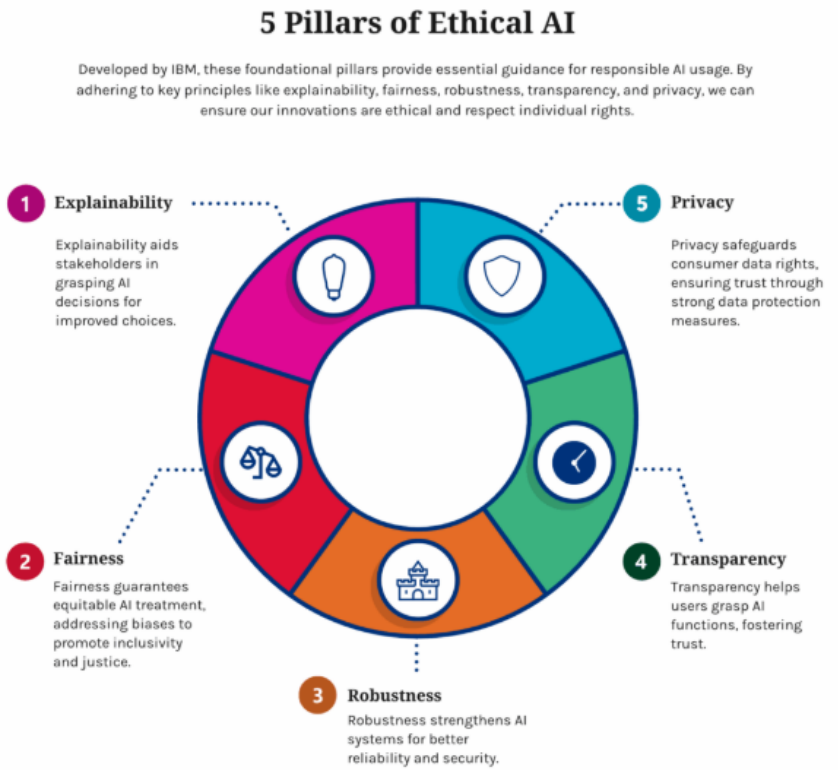


Figure 6. The overlaps between algorithm robustness, fairness, explainability and privacy.

https://www.centraleyes.com/ai-auditing/



WHAT IS ETHICS-BASED AI AUDITING?

- Auditing procedures for which voluntary ethics principles serve as the normative baseline.
- Ethics-based auditing can be either collaborative or adversarial.
- If collaborative: audits are conducted in collaboration with technology providers to assess whether their AI systems adhere to predefined ethics principles […].
- If adversarial: independent actors conduct audits to assess an AI system without access to its source code.
- Collaborative audits aim to provide assurance
- Adversarial audits aim to expose harms.

Elise Racine & The Bigger Picture / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

WHAT IS ETHICS-BASED AI AUDITING?

In both cases (adversarial or collaborative), ethics-based auditing concerns what ought to be done over and above compliance with existing regulations.

Elise Racine & The Bigger Picture / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Thank you!

Reihaneh Golpayegani & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/