



THE UNIVERSITY *of* EDINBURGH

Case Studies in Responsible Natural Language Processing – CS-NLP

Prof Frauke Zeller, Ella Markham



Week 1 – Introduction to the Course

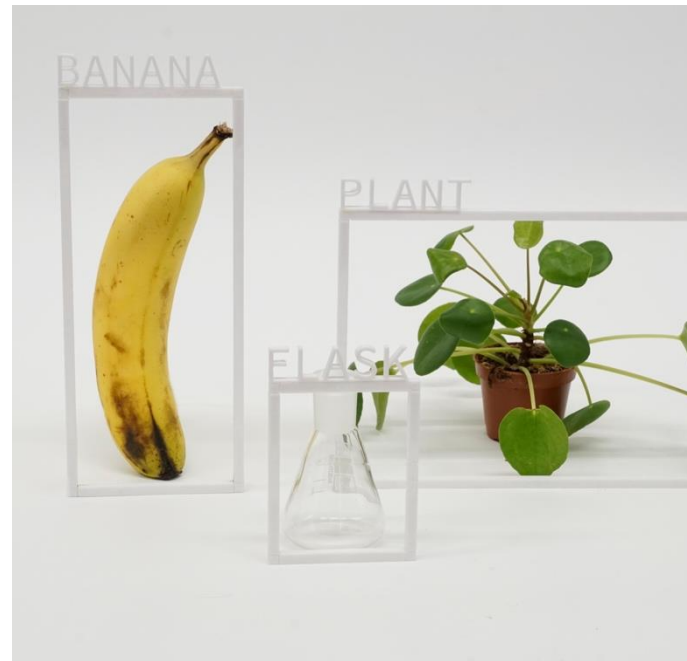
- Who are we
- What the course is about
- What do you expect, how can we work together
- Assessments
- AOB

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 progress in class on a **weekly** basis to be able to better reflect on our own disciplinary bias and arrive at new ideas.



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

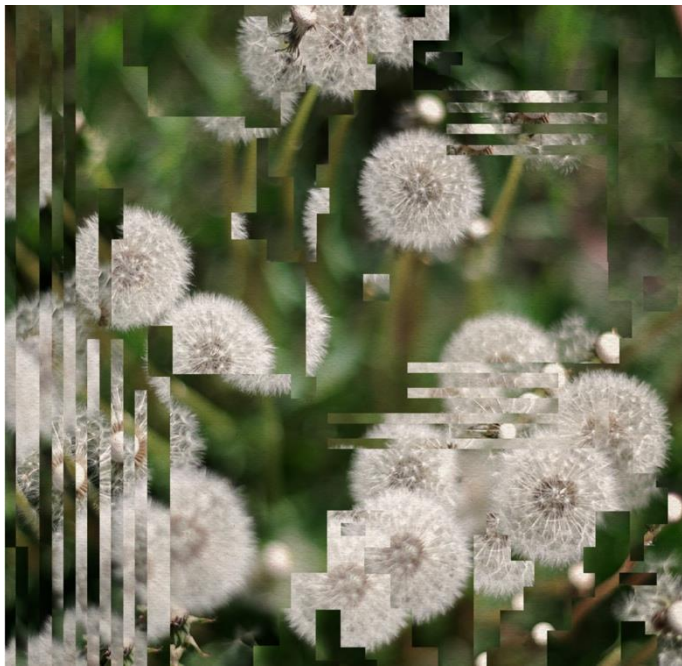
The course will focus on applying responsible NLP principles in practice. This will take two forms:

- (1) you will report and reflect on your year 1 learning ideally in the areas 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/>



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

Learning Outcomes:

On completion of this course, you 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, assessed as one:

1. Seminar presentation on a topic in responsible NLP;
2. Report on an analysis of a case study/AI Audit suggested by a CDT partner.

Both assignments are group work.



Leo Lau & Digit / <https://betterimagesofai.org/> /
<https://creativecommons.org/licenses/by/4.0/>



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 you consider vital.



Assignment 1 – the Nitty Gritty:

- You will work in groups of 4-5 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, 15-20 minutes plus 10 minutes guided discussion
- The presentation should be interactive and include class participation.
- 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:

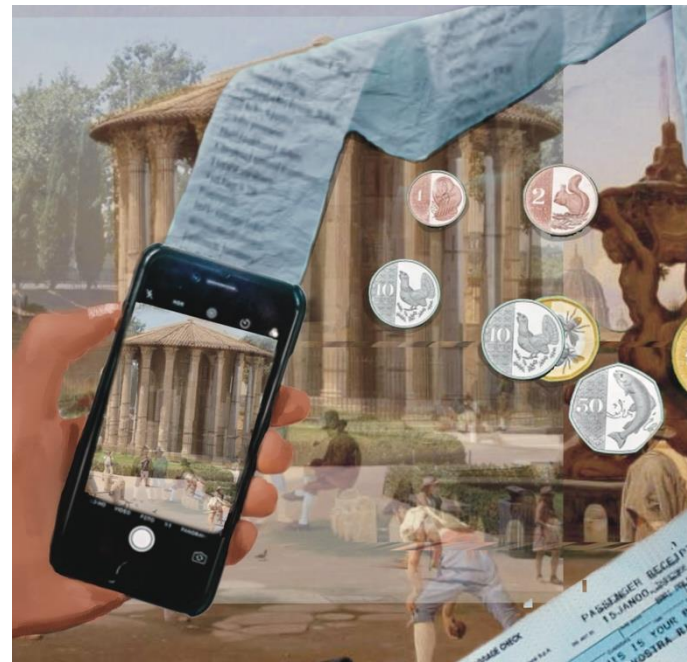
The assignment part of a larger assessment, which will be graded in one go after submission of the second assessment. Formative feedback for the presentation will be provided on:

- overall performance, time management
- pacing, voice projection, natural/free speaking
- composition and design of slides,
- content,
- discussion moderation,
- in-class work organisation and management
- overall group work management

Assignment 2:

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

- You will work in groups of 4-5
- You will choose one of the offered cases to conduct an AI “audit”.
- The terms and details of the audit will be defined together jointly with the partner, 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).
- Presentations will be around 15 minutes (in group).
- 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 (3000 words, including references) to Turnitin. Additional material will be ignored by the markers.



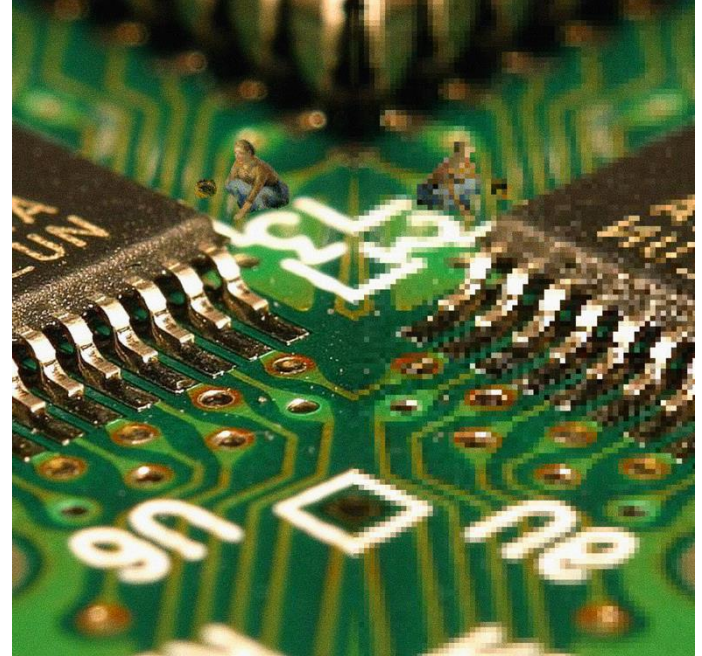
Assignment 2 – the Nitty Gritty:

Summative feedback will be provided on the submitted final report as well as the other parts of the overall assessment. The markers will provide concrete assessment criteria during class.

Formative feedback on the design of the case study, the ongoing work and the presentation will be provided during the term – in specific check-in sessions, and at the end of the term, after the final presentations.



WHAT IS AN AUDIT IN AI?



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

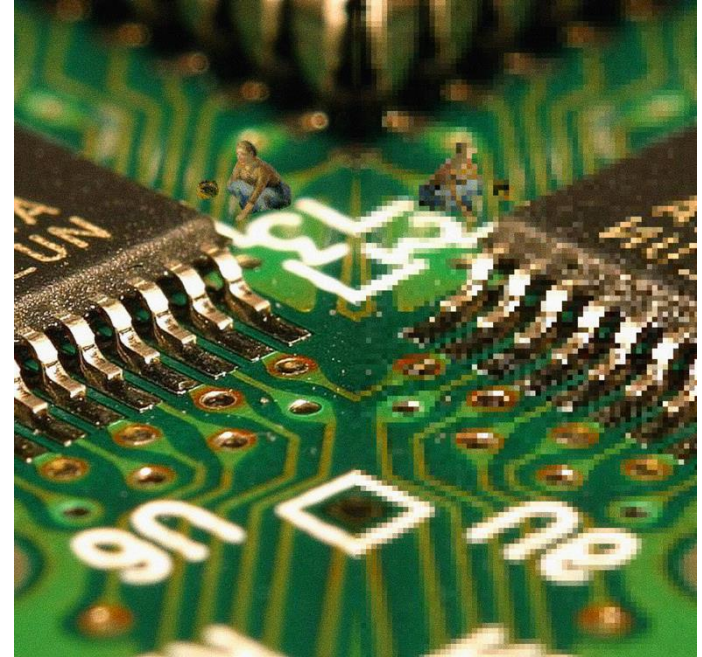
WHAT IS AN AUDIT IN AI?

1 a : a formal examination of an organization's or individual's accounts or financial situation

1 b : the final report of an audit

2 : a methodical examination and review

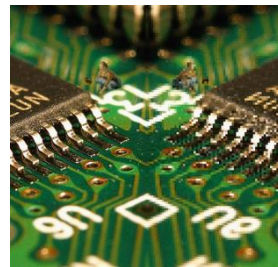
(Merriam-Webster)



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)



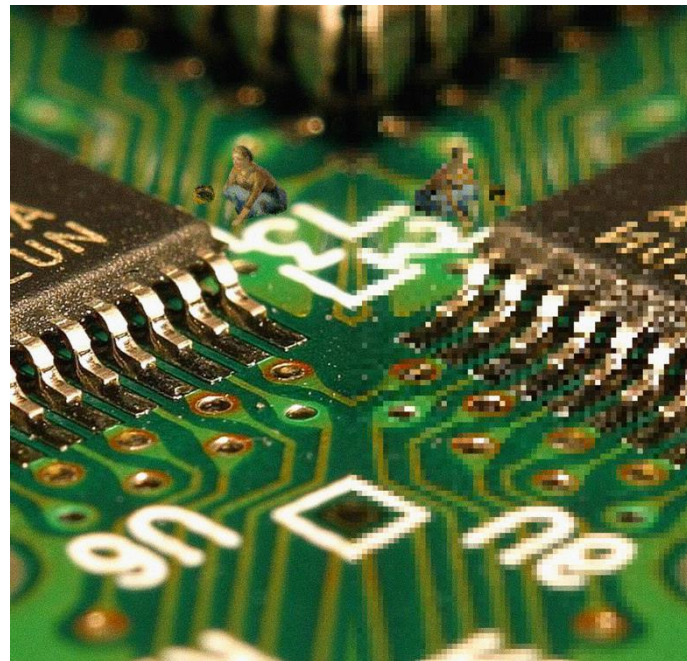
“In the social sciences, the term ‘audit study’ refers to a research method, specifically a type of field experiment, which is used to examine individuals’ behaviour or the dynamics of social processes” (Gaddis, 2018; in Mökander, 2023, p.7).

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/>

Top-down:

Regulators need new *enforcement mechanisms* to ensure that the design and use of autonomous and self-learning technologies are legal, ethical and technically robust

Corporate governance

- Internal vs external,
- Technical vs process

**Need for
auditing of AI**

Emerging legislation

- EU AIA
- US AAA

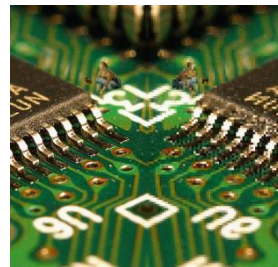
Bottom-up:

Organisations that design and deploy autonomous and self-learning technologies have incentives to implement and maintain appropriate *governance mechanisms*

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



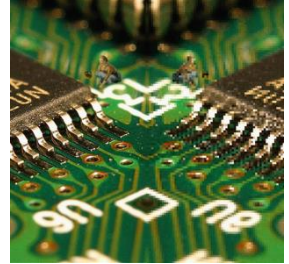
WHAT IS AI GOVERNANCE?





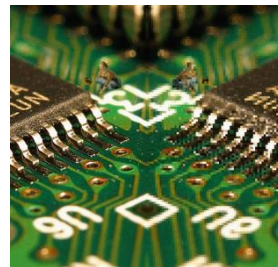
WHAT IS AI GOVERNANCE?

“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.nty.m.ki et al., 2022, p.2).”





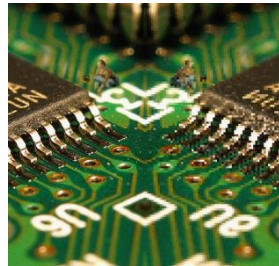
WHAT IS AN ALGORITHM AI AUDIT?



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).”



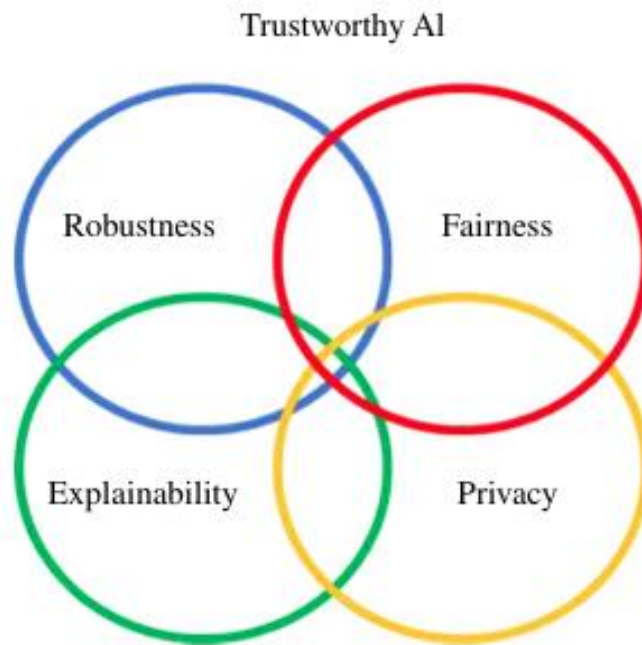
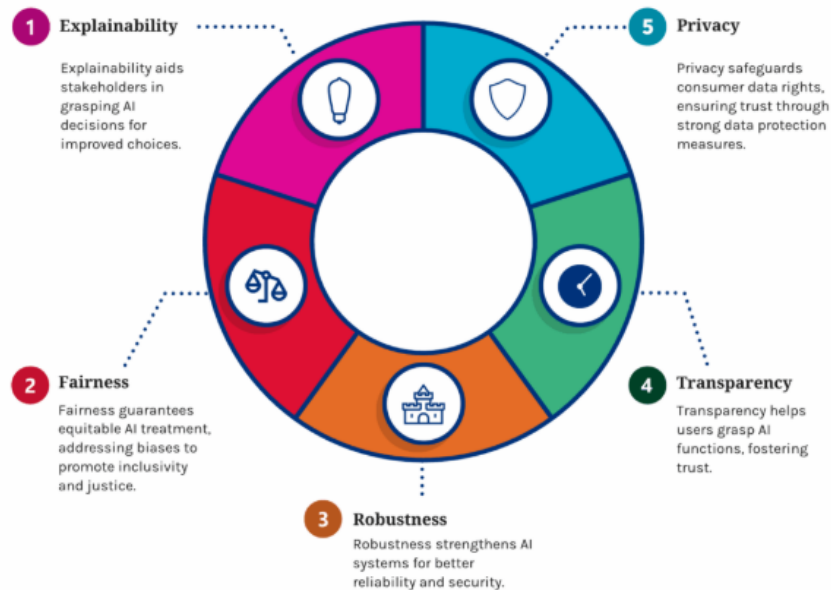


Figure 6. The overlaps between algorithm robustness, fairness, explainability and privacy.

5 Pillars of Ethical AI

Developed by IBM, these foundational pillars provide essential guidance for responsible AI usage. By adhering to key principles like explainability, fairness, robustness, transparency, and privacy, we can ensure our innovations are ethical and respect individual rights.

<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/>