Course Manual

Fairness, Accountability, Confidentiality and Transparency in AI (FACT-AI)
MSc Artificial Intelligence, January 2025
University of Amsterdam

Fernando P. Santos
f.p.santos@uva.nl

Melika Davood Zadeh
m.davoodzadeh@uva.nl

Pieter Pierrot
p.j.pierrot@uva.nl

Jacobus Martin Smit
j.m.m.smit@uva.nl

Madhura Pawar
madhura.pawar@student.uva.nl

Alexandre da Silva Pires
a.m.dasilvapires@uva.nl

Maria Heuss
m.c.heuss@uva.nl

Clara Rus
c.a.rus@uva.nl

Antonios Tragoudaras
antonios.tragoudaras@student.uva.nl

Zoe Tzifa-Kratira
zoe.tzifa.kratira@student.uva.nl

Konrad Szewczyk
konrad.szewczyk@student.uva.nl

Jesse Wonnink
jesse.wonnink@student.uva.nl

Akis Lionis
akis.lionis@student.uva.nl

Floris Six Dijkstra
floris.six.dijkstra@student.uva.nl

Roan van Blanken
roan.van.blanken@uva.nl

Rohith S. P. Prabakaran
rohith.saai.pemmasani.prabakaran@student.uva.nl

1 INTRODUCTION

The objective of this course is understanding the *technical* aspects of Fairness, Accountability, Confidentiality, Transparency in AI, specifically getting to know current challenges in the area and state-of-the-art algorithms. We will also discuss recent topics related with impacts of Generative AI on society and Environmental aspects of AI. Students will discuss influential articles in the field and work on reproducing recent approaches, thereby also gaining hands-on experience on the process of scientific research in AI and contribute to good reproducibility practices in the field.

1.1 Fairness

Research on fairness primarily involves mitigating the algorithmic discrimination of individuals based on protected attributes such as gender or race. There are many (oftentimes competing) fairness metrics resulting in a wide range of ways to frame the fairness problem and design techniques to reduce bias [4].

1.2 Accountability

Research on accountability is usually centred around identifying the actors responsible for the (potentially incorrect or unjust) algorithmic decisions. As a result, this research line is typically less focused on the algorithms themselves and instead focuses on the societal *impact* of algorithms, both in the short and long-term. Transparency and reproducibility are fundamental ingredients when conceiving techniques for accountability [3].

FACT-AI Course, January 2025, University of Amsterdam

1.3 Confidentiality

Research on confidentiality examines how the privacy of individuals whose data is being used to develop ML models can be preserved. If data is high-dimensional and there is a wide range of possible values per feature, specific individuals can be identified; anonymization by removing personal identifiers such as name or ID is not enough. We will focus on the concept of differential privacy [1].

1.4 Transparency

Research on transparency involves interpreting the behaviour of complex models. This is typically done in either a global (interpreting the whole model) or local (interpreting individual predictions) manner. In this course we will primarily focus on the latter, which involves methods such as identifying important features, generating counterfactual examples, or finding prototypical examples of a particular class [5].

1.5 Generative AI

Generative (foundation) models have recently become widely popular. These systems can generate text, images, code, or other forms of information based on an input prompt. While capable of impressive results, generative models have been also associated with risks such as proliferating low-quality data and misinformation, eroding democracy, homogenizing content, impacting negatively creative industries, and augmenting economic concentration. We will discuss these risks and mitigation strategies [2].

1.6 Environment

Current deep learning models, particularly large language models, have hundreds of billions of parameters and training them requires a lot of energy and water. After trained, these models receive billions of queries a day. The environmental impacts of large deep

learning models during training and inference are high and often unclear to developers and users. We will discuss how to quantify the environmental impact of AI and best practices to reduce it [6].

2 PROJECT DESCRIPTION

The lack of reproducibility has been an ongoing issue in academic research. The goal of the FACT-AI course project is to expose students to state-of-the-art research in the FACT field(s) while assessing the reproducibility of existing work by reimplementing an algorithm, replicating and/or extending the experiments from the corresponding paper, and detailing findings in a report.

In this project you will implement an existing FACT algorithm in groups of 4. We will follow the setup from the Machine Learning Reproducibility Challenge (MLRC) 2025 (`https://reproml.org`). We encourage you to participate in the challenge by submitting your work to the Transactions on Machine Learning Research (TMLR). The challenge task description specifies: *"Essentially, think of your role as an inspector verifying the validity of the experimental results and conclusions of the paper. In some instances, your role will also extend to helping the authors improve the quality of their work and paper."* More details on participating in the MLRC 2025 here: `https://reproml.org/call_for_papers/`.

There are two scenarios possible for this project:

(1) There already exists an open-source implementation of your selected paper. You are allowed to use this, but we will be aware of the fact that this implementation is available. Given the implementation:
   (a) The results you obtain are different as described in the paper (i.e. the paper is not reproducible). Your report should explain what these differences are and conjecture why they occur. You should also try to resolve the problem(s) and explain your rationale behind the choices you made, as well as describing your implementation process and the results you obtained.
   (b) The results are reproducible, meaning this method can now be used for further research. The experimental results are less robust when they do not scale beyond the original model, data(s) and domain(s) used in the paper. Are these results also reproducible for other domains, datasets, model (configurations), etc?
(2) There is no open-source implementation available, meaning your group needs to re-implement everything from scratch. What are the difficulties while reproducing this work and how have you solved them? Is there enough information in the paper to reproduce the results? Are the results similar as described in the paper? If not, why? If yes, is this work reproducible for other domains, datasets, configurations?

If an open-source implementation exists, the result 'the paper is reproducible' is not enough for a good grade. Either you need to go beyond the original results by questioning the results on other domains, data, and/or model configurations, or you need to show that the results are not as in the paper and propose an alternative solution. Note that the final submission of your implementation must be done in **PyTorch/Python**, which might not necessarily be the language of the available code. If there is no open-source implementation, the report should explain in detail how and if the work is reproducible. Please, always keep in mind what is described in the original MLRC task: *The goal of this challenge is not to criticize papers or the hard work of our fellow researchers. Science is not a competitive sport. Thus, the main objective of this challenge is to enable a mutually beneficial learning experience, while contributing to the research by strengthening the quality of the original paper.*

The deadline for handing in the project on Canvas is **23:59 on 31 January 2025**. Please note that the deadline to submit to the challenge and declare your intent is three weeks after the course finishes (February 21). Participating in the TMLR and the challenge is a great opportunity to understand how ML research is done by interacting with reviewers and getting feedback on your work. If your paper gets accepted in the challenge, you might be able to participate in conference at Princeton University (more details: `https://canvas.uva.nl/courses/45955/pages/papers-to-reproduce/`).

2.1 Report

To participate in the challenge, you must follow the instructions in `https://reproml.org/call_for_papers/`). Note that although MLRC includes all papers from top conferences, we are only focusing on papers about FACT topics. You can find a list of potential papers to select in Canvas: `https://canvas.uva.nl/courses/39312/pages/papers-to-reproduce`

To write the report, you should use the TMRL latex template and follow the Submission Guidelines and Editorial Policies of this journal: `https://jmlr.org/tmlr/editorial-policies.html`.

The objective of the report is to explain the results you obtained as well as the process behind the implementation. Your report should be **no more than 10 pages long (excluding references)**.

If you would like to receive feedback on an early draft of your report (which is highly recommended) you can email it to your TA by **23:59 on 23 January 2025**. You will need to submit the final report via Canvas by **23:59 on 31 January 2025**.

2.2 Final Code Submission

The final submission of your implementation should be in a private GitHub repository with all the information, code and data needed to test your implementation. Any commits you make to your repository after the deadline will be ignored. All implementations requiring a deep learning framework **must be done in PyTorch**. Please set your repository up in a clean and reasonable way with the following components:

- Environment configuration.
- IPython (Jupyter) notebook detailing all results in the report. Please ensure that it is possible to run all cells and obtain the results without any issues. Make sure that only the code for generating the results is present in the notebook. The model(s) and all the other files needed to generate the results should be in separated files.
- Instructions for how to run your implementation.
- Dataset(s) used in the experiments.
- All required scripts for testing the implementation.

Table 1: Lecture schedule for the course. Lectures will occur onsite, at Science Park. Please check DataNose for room numbers.

	Mon 6 Jan	**Wed 8 Jan**	**Fri 10 Jan**	**Mon 13 Jan**	**Mon 20 Jan**	**Mon 27 Jan**
11:00	Introduction to the course	Transparency lecture	Privacy lecture	Fairness lecture	Accountability lecture	Generative AI lecture
12:00	Environment and Resources lecture	Guest lecture Ana Lucic	Guest lecture Rob Romijnders	Guest lecture Ulle Endriss	Guest lecture Vanja and Mirthe	Guest lecture Nanne van Noord
-	**Deadline to form groups**	-	Quiz 1	Quiz 2	Quiz 3	Quiz 4

We also want your code to be reproducible. Please take a look at the following resources for suggestions and best practices on producing reproducible code: (1) `https://github.com/paperswithcode/releasing-research-code`, and (2) `https://www.cs.mcgill.ca/~ksinha4/practices_for_reproducibility/` .

2.3 Presentation

The final part of the project is a 10 minute presentation on your findings. This should be a summary of your written report and will take place during the last week of the course, on 31 January 2025 (exact times to be scheduled, any time from 9:00 to 15:00). Presentations will happen onsite at Science Park.

2.4 Grading

A Grading Matrix will be provided in the first week of the course (see Canvas). Regarding the submission to TMLR and the MLRC challenge: please keep in mind that if you submit the paper will be publicly available on OpenReview and therefore anyone can see it. You should only submit after discussing the quality of your report with your TA supervisor.

The Introduction section of your report should establish a bridge with the high-level FACT topics discussed in the first week of the course, an aspect that we will pay special attention to when grading your work.

3 LOGISTICS

Please complete the following steps **by 2:59PM on 6 January**:

(1) Choose your group for the project. There should be a maximum of 4 students per group. All communication about the project should take place with the entire group.
(2) Discuss with your group which papers you would like to implement from the list of papers provided in Canvas.
(3) Create a private GitHub repository for your project. All communication will be handled (and logged) via issues in this repository.
(4) **One person per group** needs to fill out the preference Google Form (see Canvas for the link). We will do our best to take everyone's paper preferences into account but given the number of students taking this course, we cannot guarantee that you will be assigned one of your top papers. The final paper, TA and Practicum assignment will be communicated in the afternoon of **January 6**.

You will have two onsite Practicums per week where you can ask your TA questions about your paper. You need to go to the Practicums that correspond to your paper. Each group will get 2×20 minute online Practicums each week. The final TA and Practicum time (which will depend on the paper you will be assigned to) will be communicated on **January 6**. Please check DataNose for specific times and rooms.

4 LECTURES

There are six lecture blocks for this course, one for each topic (Privacy, Transparency, Fairness, Accountability, Generative AI and Environment). The schedule can be found in Table 1. Each lecture block will consist of (i) a general lecture on the topic, (ii) guest lecture by an invited speaker on the same topic. We are very happy to have a set of distinguished guest speakers participating in our lectures:

- The transparency guest lecture (Jan 10, 12:00-12:45) will be led by **Ana Lucic**, former PhD student at UvA and former coordinator of the FACT-AI course. Ana completed her PhD in Explainable AI, moved to Microsoft Research for a research position, and is currently an Assistant Professor of Explainable AI at IvI/ILLC.
- The Privacy guest lecture (Jan 10, 12:00-12:45) will be led by **Rob Romijnders**, PhD student at IvI / Amsterdam Machine Learning Lab (AMLab). Rob is doing his PhD in representation learning, decentralized machine learning, and differential privacy.
- The Fairness guest lecture (Jan 13, 12:00-12:45) will be led by **Ulle Endriss**, Professor of AI and Collective Decision Making at the Institute for Logic, Language and Computation (ILLC), working on problems at the interface of AI, economics and political science (computational social choice).
- The Accountability guest lecture (Jan 20, 10:00-10:45) will be led by **Vanja Škorić**, Program Director at the Netherlands-based European Center for Not-for-profit Law, and **Mirthe Dankloff**, PhD candidate at the Civic AI Lab and the User centric data science group at the VU.
- The Generative AI guest lecture (Jan 27, 10:00-10:45) will be led by **Nanne van Noord**, Assistant Professor at the Multimedia Analytics Lab of the University of Amsterdam and an expert in Multimedia Analysis with and for Visual Culture.

5 QUIZZES

The contents taught in the lectures will be exercised an evaluated in a set of quizzes. These quizzes will be done through Canvas and will consist of approximately 15 multiple choice questions to be answered in 15 minutes. The total weight of all quizzes on the final grade will be 15%. There will be 4 quizzes:

- January 10: Quiz on Confidentiality, Environment and Transparency (7.5%)
- January 13: Quiz on Fairness (2.5%)
- January 20: Quiz on Accountability (2.5%)
- January 27: Quiz on Generative AI (2.5%)

6 PAPERS TO BE REPRODUCED

You will implement **one** of the 20 possible papers listed in Canvas with your group. We expect a **maximum of 3-4 groups** working on the same paper. Please check the list of possible papers via Canvas (`https://canvas.uva.nl/courses/45955/pages/papers-to-reproduce`).

REFERENCES

[1] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
[2] S. Huang and D. Siddarth. Generative ai and the digital commons. *arXiv preprint arXiv:2303.11074*, 2023.
[3] B. Kim and F. Doshi-Velez. Machine learning techniques for accountability. *AI Magazine*, 42(1):47–52, 2021.
[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.
[5] C. Molnar. *Interpretable machine learning*. 2020.
[6] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.