Course Manual

Fairness, Accountability, Confidentiality and Transparency in AI (FACT-AI)
MSc Artificial Intelligence, January 2022
University of Amsterdam

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

Maurits Bleeker
m.j.r.bleeker@uva.nl

Sami Jullien
s.jullien@uva.nl

Ana Lucic
a.lucic@uva.nl

Ilse van der Linden
i.w.c.vanderlinden@uva.nl

Sara Altamirano
s.e.altamirano@uva.nl

Adeel Pervez
a.a.pervez@uva.nl

Wenzhe Yin
w.yin@uva.nl

Tim Alpherts
t.o.l.alpherts@uva.nl

Mayesha Tasnim
m.tasnim@uva.nl

Dimitris Michailidis
d.michailidis@uva.nl

Christos Athanasiadis
c.athanasiadis@uva.nl

Yangjun Zhang
y.zhang6@uva.nl

1 INTRODUCTION

The objective of this course is understanding the *technical* aspects of each of the four topics (Fairness, Accountability, Confidentiality, Transparency), specifically existing algorithms.

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 different (oftentimes competing) fairness definitions resulting in a wide range of ways to frame the problem.

1.2 Accountability

Research on accountability is usually centred around identifying who is responsible for the (potentially incorrect or unjust) decision that is a result of an algorithmic prediction. As a result, this research is typically less focused on the algorithms themselves and instead places the focus on the *impact* of algorithms.

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 the data is high-dimensional and there is a wide range of possible values per feature, the information is essentially identifiable, and therefore simple anonymization of personal identifiers such as name or email is not enough.

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.

2 PROJECT DESCRIPTION

The lack of reproducibility has been an ongoing issue in academic research. The goal of this FACT-AI course project is to assess the reproducibility of existing work by reimplementing an algorithm, replicating and/or extending the experiments from the corresponding paper, and detailing your findings in a report. In this assignment you will implement an existing FACT algorithm in groups of 4. We will follow the setup from the Machine Learning Reproducibility Challenge (MLRC) (`https://paperswithcode.com/rc2021`), and encourage you to participate in the challenge by submitting your work (submission deadline: **February 4, 23:59**). The 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."* The full task description is available here: `https://paperswithcode.com/rc2021/task`.

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 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

FACT-AI Course, January 2022, University of Amsterdam 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 reimplement everything yourselves. 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 is reproducible for other domains, datasets, model (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. The deadline for handing in the project on Canvas is **23:59 on 4 February 2022**, which is also the same as the deadline for submitting to the MLRC. This is intentional – formally participating in the challenge is a great opportunity to understand how ML research is done by interacting with reviewers and getting feedback on your work.

2.1 Report

To participate in the challenge, you must follow the instructions in `https://paperswithcode.com/rc2021/registration`). Note that although MLRC includes all papers from top conferences, we are only focusing on papers about FACT topics. All papers suggested below (section 4) qualify to the MLRC.

To write the report, you will use the MLRC template: `https://www.overleaf.com/project/5f4e72de7681920001b208f9`. 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 8 pages long (excluding references)**.

If you would like to receive feedback on an early draft of your report, you can email it to your TA by **23:59 on 27 January 2022**. You will need to submit the final report via Canvas by **23:59 on 4 February 2022**.

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 simply 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 needs to be generate the results should be in separated files. It should function as some kind of API.
- Instructions for how to run your implementation.
- Dataset(s) used in the experiments.
- All required scripts for testing the implementation.

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 essentially be a summary of your written report and will take place during the last week of the course, on 4 February 2022 (exact times to be scheduled, any time from 9:00 to 17:00).

2.4 Grading

A Grading Matrix will be provided in the first week of the course (see Canvas). If you submit to the MLRC, you will get an extra 0.5 point. Please keep in mind that if you submit, the paper will be publicly available on OpenReview and therefore anyone can see it. We will also award another 0.5 point the 10 groups with the best code implementations.

3 LOGISTICS

Please complete the following steps **by 23:59 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 Section 4.
(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 following Google Form: `https://forms.gle/yaccMt5NFHV9XkvA6` We will do our best to take everyone's paper preferences into account but given the number of students taking this course, we simply cannot guarantee that you will be assigned one of your top papers.

Each group will get 2 × 20 minute online Practicums each week.

4 PAPERS TO BE REPRODUCED

You will implement **one** of the following papers with your group. We expect a **maximum of 3 groups** working on the same paper.

- W. Gao, S. Guo, T. Zhang, H. Qiu, Y. Wen, and Y. Liu. Privacy-preserving collaborative learning with automatic transformation search. In *CVPR'21*, pages 114–123, 2021
- O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al.

Table 1: Lecture schedule for the course. Due to COVID-19, all lectures will be online via Zoom.

	Mon 10 Jan	Wed 12 Jan	Fri 14 Jan	Mon 31 Jan
9:00 - 10:00	Intro and Reproducibility lecture	Fairness (Guest) lecture	Transparency lecture	Accountability lecture
10:00 - 11:00	Confidentiality lecture	Guest lecture	Discussion group #3	Guest lecture
11:00 - 12:00	Discussion group #1	Discussion group #2	Paper dissection #3	
12:00 - 13:00	Paper dissection #1	Paper dissection #2	Guest lecture	

Explaining in style: Training a gan to explain a classifier in stylespace. *ICCV'21*, 2021

- J. Correa, A. Cristi, P. Duetting, and A. Norouzi-Fard. Fairness and bias in online selection. In *ICML'21*, pages 2112–2121. PMLR, 2021
- Y. Choi, M. Dang, and G. Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. *AAAI'21*, 2021
- B. van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *NeurIPS'21*, 34, 2021
- M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, and Y. Zhang. Robust counterfactual explanations on graph neural networks. *NeurIPS'21*, 34, 2021
- L. Li, B. Wang, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *ICCV'21*, pages 1046–1055, 2021
- A. Sauer and A. Geiger. Counterfactual generative networks. *ICLR'21*, 2021
- N. Raman, S. Shah, and J. Dickerson. Data-driven methods for balancing fairness and efficiency in ride-pooling. *IJCAI'21*, 2021
- J. K. Lee, Y. Bu, D. Rajan, P. Sattigeri, R. Panda, S. Das, and G. W. Wornell. Fair selective classification via sufficiency. In *ICML'21*, pages 6076–6086. PMLR, 2021
- N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan. Exacerbating algorithmic bias through fairness attacks. *AAAI'21*, 2021
- I. M. Ziko, E. Granger, J. Yuan, and I. B. Ayed. Variational fair clustering. *AAAI'21*, 2021
- S. Upadhyay, S. Joshi, and H. Lakkaraju. Towards robust and reliable algorithmic recourse. *NeurIPS'21*, 2021
- S. Levanon and N. Rosenfeld. Strategic classification made practical. *ICML'21*, 2021
- M. Antoniak and D. Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *ACL'21*, pages 1889–1904, 2021
- Y. Du, Q. Fang, and D. Nguyen. Assessing the reliability of word embedding gender bias measures. *EMNLP'21*, 2021

5 LECTURES

There are four lecture blocks for this course, one for each topic. The schedule can be found in Table 1. Each lecture block will consist of (i) a general lecture on the topic, (ii) a student discussion group on a prominent paper, and (iii) a "paper dissection" session, where a TA will go over the same prominent paper. The purpose of the discussion group is to learn how to pick apart research papers, since this is an important part of reimplementing existing work (and an important part of research in general). **To prepare for the discussion group, you are expected to read the paper in advance and contribute to the discussion.** In the corresponding paper dissection session, a TA will go over the same paper you discussed in the discussion group, to give an overview of the papers' strengths and weaknesses.

We will also have four guest lectures:

- The background on Fairness (guest) lecture will be by Emma Beauxis-Aussalet, Assistant Professor at the VU and Lab Manager at the Civic AI Lab, doing research on the practical means to harness the impacts of AI on society.
- The Fairness guest lecture will be by Ulle Endriss, Professor of Artificial Intelligence 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 Transparency guest lecture will be by Violeta Misheva, an expert in Explainable AI working as a data scientist at ABN AMRO bank in Amsterdam.
- A final guest lecture will be given by Yuki M. Asano, Assistant Professor for computer vision and machine learning at the QUVA lab at the University of Amsterdam, also working on privacy and bias in Machine Learning.

5.1 Papers for Dissections

The paper dissection sessions will be led members of the IRLab and the Civic AI Lab at the UvA. We will cover the following papers (please read the in advance and prepare discussion points or questions):

- **Paper dissection #1 - Confidentiality:** M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC 2016*, 2016. `https://arxiv.org/pdf/1607.00133.pdf`. This session will be led by Sami Jullien.
- **Paper dissection #2 - Fairness:** L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed Impact of Fair Machine Learning. In *ICML 2018*. This session will be led by Fernando Santos. `https://arxiv.org/abs/1803.04383`
- **Paper dissection #3 - Transparency:** M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD 2016*, 2016. `https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf`. This session will be led by Ilse van der Linden.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC 2016*, 2016.

[2] M. Antoniak and D. Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *ACL'21*, pages 1889–1904, 2021.

[3] M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, and Y. Zhang. Robust counterfactual explanations on graph neural networks. *NeurIPS'21*, 34, 2021.

[4] Y. Choi, M. Dang, and G. Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. *AAAI'21*, 2021.

[5] J. Correa, A. Cristi, P. Duetting, and A. Norouzi-Fard. Fairness and bias in online selection. In *ICML'21*, pages 2112–2121. PMLR, 2021.

[6] Y. Du, Q. Fang, and D. Nguyen. Assessing the reliability of word embedding gender bias measures. *EMNLP'21*, 2021.

[7] W. Gao, S. Guo, T. Zhang, H. Qiu, Y. Wen, and Y. Liu. Privacy-preserving collaborative learning with automatic transformation search. In *CVPR'21*, pages 114–123, 2021.

[8] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. *ICCV'21*, 2021.

[9] J. K. Lee, Y. Bu, D. Rajan, P. Sattigeri, R. Panda, S. Das, and G. W. Wornell. Fair selective classification via sufficiency. In *ICML'21*, pages 6076–6086. PMLR, 2021.

[10] S. Levanon and N. Rosenfeld. Strategic classification made practical. *ICML'21*, 2021.

[11] L. Li, B. Wang, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *ICCV'21*, pages 1046–1055, 2021.

[12] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed Impact of Fair Machine Learning. In *ICML 2018*.

[13] N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan. Exacerbating algorithmic bias through fairness attacks. *AAAI'21*, 2021.

[14] N. Raman, S. Shah, and J. Dickerson. Data-driven methods for balancing fairness and efficiency in ride-pooling. *IJCAI'21*, 2021.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD 2016*, 2016.

[16] A. Sauer and A. Geiger. Counterfactual generative networks. *ICLR'21*, 2021.

[17] S. Upadhyay, S. Joshi, and H. Lakkaraju. Towards robust and reliable algorithmic recourse. *NeurIPS'21*, 2021.

[18] B. van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *NeurIPS'21*, 34, 2021.

[19] I. M. Ziko, E. Granger, J. Yuan, and I. B. Ayed. Variational fair clustering. *AAAI'21*, 2021.