Seminar Description, SS20

Ethics in AI

Florian Pfisterer, Susanne Dandl, Gunnar König, Christoph Molnar, Timo Freiesleben
{florian.pfisterer, susanne.dandl, gunnar.koenig,
christoph.molnar}@stat.uni-muenchen.de
timo.freiesleben@campus.lmu.de

Course Description

Automation of decision processes is becoming increasingly ubiquitous in a digital era and increasingly affects human lives. In this seminar, we want to raise awareness for several important issues in the intersections of data science, philosophy and law. Topics discussed in the seminar range from acquiring and storing data to problems in the area of fairness, accountability and transparency (FACT). The seminar will try to discuss questions in the aforementioned areas from a technical, philosophical and judicial perspective, trying to create a set of best practices for data handlers, data scientists and decision makers.

General Information

Kickoff and Introduction: 03.11.2020, 4pm CEST

Discussion of Block 1: 18.12.2020, 1pm - 4pm CEST

Discussion of Block 2: 29.01.2020, 1pm - 4pm CEST

Every topic either belongs to block 1 or block 2. Every presenter from block 1 is randomly assigned to one sparring partner from block 2 and vice versa. Consequently, for every topic there is an author and a reviewer.
We move the focus from the essay/paper to the presentation and the interaction within the group. As the work done as a Sparring partner counts as “Koreferat”, shorter essays are admitted by the “Prüfungsordnung”.
Zoom Links will be made available on mattermost and in moodle shortly before the meeting.

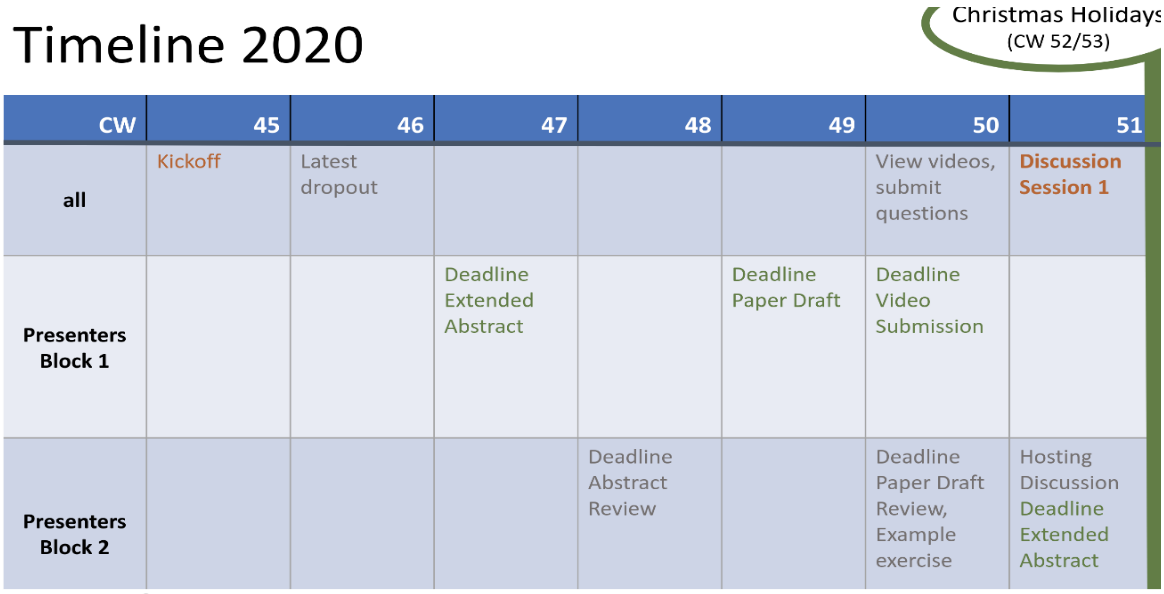
Mattermost Channel: [mattermost link] Join “*seminar_ethics_in_ai*” Mattermost channel.
Moodle: [moodle link], Key: *ethicsAI_20*

Overview deliverables (and deadlines) by role:

- Author:
  - Extended abstract (4 weeks before the Discussion Block, until Friday 23:59 pm via Mattermost channel)
  - Paper draft (2 weeks before the Discussion Block, until Friday 23:59 pm via Mattermost channel)

    - [30 Min] Presentation video & slides as pdf (1 week before the Discussion Block, until Friday 23:59 pm via Mattermost channel)
    - [5 Min] Pitch video (until day of Discussion Block via Mattermost channel)
    - Final essay (2 weeks after the Discussion Block, until Friday 23:59 pm via Mattermost channel)
    - Folder to compile presentation slides & pitch video (2 weeks after the Discussion Block, until Friday 23:59 pm via PM or email to supervisor)
- Sparring partner:
    - Review Extended Abstract (1 week after receival, until Friday 23:59 pm, via PM or email to author and supervisor of respective topic): Is the overall structure of the paper clear? Is anything missing?
    - Review Paper Draft (1 week after receival, until Friday 23:59 pm, via PM or email to author and supervisor of respective topic): Is the structure and content of the paper understandable? The review will be in plain text and is sent to the author and supervisor.
    - [Optional] Create a small exercise sheet for your colleagues to test their knowledge on the topic or to fuel the discussion, e.g., a short questionnaire.
    - Host discussion: Explain your position [5 min], collect questions and lead the discussion [25 min], talk about exercise (optional), keep time

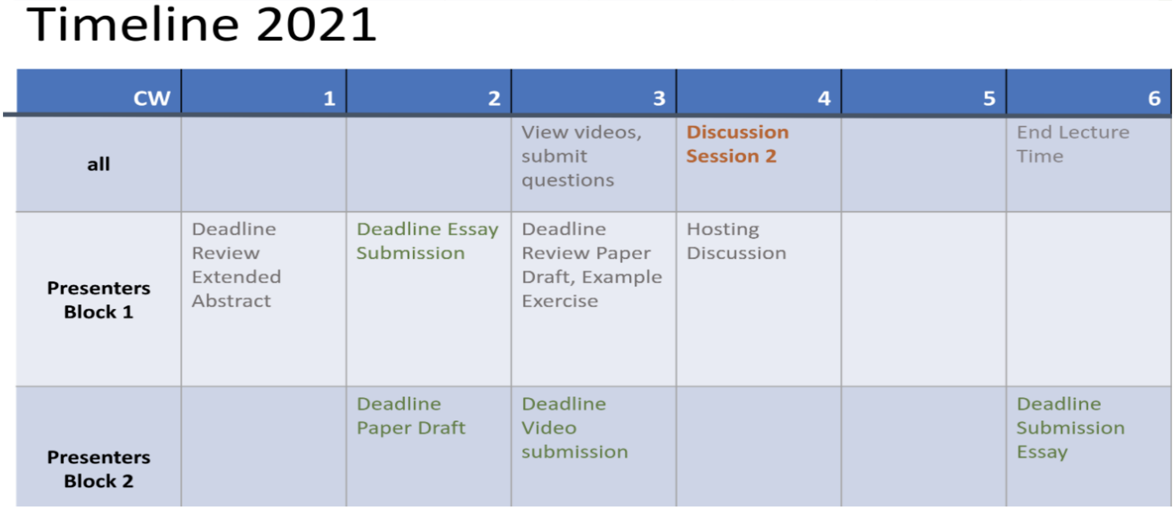
- Participant:
    - Active participation & discussion
    - Prepare at least one question for each video
    - Try to solve exemplary exercise (if provided)



Timeline 2020

Christmas Holidays
(CW 52/53)

CW	45	46	47	48	49	50	51
all	Kickoff	Latest dropout				View videos, submit questions	**Discussion Session 1**
Presenters Block 1			Deadline Extended Abstract		Deadline Paper Draft	Deadline Video Submission	
Presenters Block 2				Deadline Abstract Review		Deadline Paper Draft Review, Example exercise	Hosting Discussion Deadline Extended Abstract



Timeline 2021

CW	1	2	3	4	5	6
all			View videos, submit questions	**Discussion Session 2**		End Lecture Time
Presenters Block 1	Deadline Review Extended Abstract	Deadline Essay Submission	Deadline Review Paper Draft, Example Exercise	Hosting Discussion		
Presenters Block 2		Deadline Paper Draft	Deadline Video submission			Deadline Submission Essay

Grading & Attendance

At the kickoff event, we will provide a high-level overview of the seminar and introduce the topics we will cover. We will then assign a topic to every one of you, while trying to accommodate your preferences. The seminar will be organized in two blocks. In every block a number of presentations will be held, followed by a general critique of the presentation and a discussion of the scientific content. Your seminar grade is determined by four components:

1. **The scientific presentation (40%)**
   The goal of your presentation is to teach your topic to your fellow students. The presentation is divided into two parts: A 30 min presentation and a 5 min pitch presentation. The slideset of both presentations should optionally be created with R Markdown and should follow the template in this github repository. The submission should include the complete folder. The presentation will be graded along two dimensions:

- Clarity (70%)
- Presentation style (30%)

Because of the current COVID-19 situation we cannot meet in person. Both presentations shall therefore be recorded and uploaded by the participants, for the other students to watch. The main presentation video will be cut at 40 minutes sharp, the pitch video at 7 min.

For each topic, we ask Students to consider and present (c.a. 1 Slide each):

- Judicial Perspective:
  Discuss the topic with respect to common data laws: GDPR, BDSG, Informed Consent and practical implications for Data Scientists [1]
- Ethical Perspective:
  Present the problem in an ethical context and discuss implications of the different ethical approaches
- Technical Perspective & Existing Solutions
  Please briefly describe existing technical hurdles and solutions

[1] T GOVERNANCE PRIVACY TEAM. (2017). EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide - Second edition. Ely, Cambridgeshire, http://www.jstor.org/stable/j.ctt1trkk7x

Students for the 3 ECTS version should prepare a 10 min presentation for the second discussion block as a summary of the presentations of the first discussion block.

2. **An Essay (30%)**
   The goal of the essay is to write a review of the topic you have been assigned to. This essay has to be submitted as PDF to your supervisor. The length of your essay depends on the number of ECTS points you receive for this seminar. If you are a Bachelor student (6 ECTS points), your essay should be 8-12 pages (ca. 10k characters). If you are a Master student (9 ECTS), your essay should be at least 10-16 pages long (ca. 15k characters). Students for the 3 ECTS version submit a short essay (4-6 pages).
   For 6 and 9 ECTS, a shortened version of the extended abstract should be included at the beginning of the manuscript which is also part of the character count.

3. **Extended abstract and first draft (10%)**
   Extended abstract: The goal of the extended abstract is to summarize motivation, key arguments and if applicable results of the paper or topic. The extended abstracts should be sent to all course participants. The respective reviewer will give feedback on the overall direction and moderate the discussion after the talk.
   The extended abstract shall be between 300 and 1200 words long.
   First draft: Submitting the draft to your reviewer will give you the chance to get feedback on your final essay. It should roughly have the structure and contents of the final submission and can be in the form of bullet points, but should be understandable by the sparring partner.

4. **Feedback, Session Chair and Active Participation (20%)**
   The reviews should be given in time and feedback for the respective document should be correct and constructive.

Review Extended Abstract: Is the overall structure of the paper clear? Is anything missing? The Review will be in plain text and is sent to the author and the supervisor.
Review First Draft: Is the structure and content of the paper clear and understandable? The review will be in plain text and is sent to the author and supervisor.
We expect the reviewer to have understood important concepts enough to encourage a discussion of the topic at hand. As session chair he/she will introduce, host and moderate the discussion for the respective session as well as managing time.

Attendance is mandatory for both, discussion block 1 and discussion block 2. If you miss a block or your own presentation, you need to provide a medical certificate and reschedule. If you fail to do either of the two, you fail the seminar. You also fail the seminar if you drop out later than one week after the kickoff.

Topics

Block 1

Topic 1: Data Security Storing & Encryption

Topic 2: What is bias?

Topic 3: Model Cards and Datasheets

Topic 4: Responsibility, Accountability & Contestability

Topic 5: Level of Autonomy

Block 2

Topic 6: Strategic Classification

Topic 7: Model Safety, Uncertainty & Robustness

Topic 8: Deep Fakes

Topic 9: Data re-extraction from Models (De-) Identification, Differential Privacy

Topic 10: Algorithmic Fairness - Technical vs. Societal Solutions

Details

Topic 1: Data Security, Storing & Encryption [BSc] (Susanne)

Due to advances in computational power, new forms of data collection (IoT/smartphones/health devices) and the increasing appreciation of data by companies in conjunction with data privacy laws (GDPR, BDSG), secure storage and distribution of data became a crucial task for companies. The student should present (1) a definition of sensitive data, (2) legal fundamentals for data security especially in the GDPR (specifically data economy), (3) implications for data scientist to securely store and distribute sensitive data (access control & encryption of data) and (4) common methods, tools or R packages to deploy these principles.

Good starting points are:
Jason Andress (2011). The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice.
http://index-of.es/Hack/The.Basics.of.Information.Security_Understanding.the.Fundamentals.pdf

Vacca, John R. Computer and Information Security Handbook, edited by John R. Vacca, Elsevier Science & Technology, 2013. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/ub-lmu/detail.action?docID=1195617. [Fulltext available via OPAC]

Topic 2: What is bias? [BSc, MSc] (Gunnar)

The topic of bias is omnipresent in the literature and media discussion concerning the ethical use of AI technology. However, seldomly bias is explicitly defined. Is bias per se negative? When is it problematic? How do different disciplines define it?
Where is bias coming from? Does it stem from the model building, the data collection or society?

A good starting point is: *Danks, David, and Alex John London. "Algorithmic Bias in Autonomous Systems." IJCAI. 2017.*

Topic 3: Model Cards and Datasheets (Christoph) [Bsc]

Models and datasets are often used in a way that was not initially foreseen and they come with limitations.
Model cards (https://arxiv.org/abs/1810.03993) are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex).
Datasheets (https://arxiv.org/abs/1803.09010) serve a similar purpose for datasets that documents their motivation, composition, collection process, recommended uses, and so on.

Topic 4: Legal requirements for XAI [BSc] (Gunnar)
Accountability & Contestability

Many high-performing AI systems are hard or impossible to understand by human stakeholders. They are also referred to as “black-boxes”. As a consequence, in Machine Learning and AI a vibrant community researching “interpretability” and “explainable AI” (XAI) has emerged. In parallel, new legal frameworks like the GDPR have been developed.
The kinds of explanations that XAI provides are diverse. It is often unclear what legal requirements algorithms should fulfil.
In this work the legal perspective shall be taken to elaborate the requirements of legal frameworks (especially the GDPR) towards XAI methods. These e.g. include accountability and contestability.

Good starting points are:
Goodman, Bryce & Flaxman, Seth. (2016). EU regulations on algorithmic decision-making and a "right to explanation". AI Magazine. 38. 10.1609/aimag.v38i3.2741.

Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law, Volume 7, Issue 2, May 2017, Pages 76–99, https://doi.org/10.1093/idpl/ipx005
Almada, Marco. "Human intervention in automated decision-making: Toward the construction of contestable systems." Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. 2019. https://dl.acm.org/doi/abs/10.1145/3322640.3326699

Topic 5: Level of Autonomy [Bsc, Mcs] (Timo)

Not only are more and more tasks dedicated to automated systems, but also is the degree of human supervision in these tasks declining. This increase in the autonomy of artificial decision-making systems indeed implies technical difficulties, but more severely it poses major legal, ethical, and conceptual challenges.
(1) the student should introduce the concepts of autonomy in AI and compare them to the corresponding concept in Law and Ethics. (2) the student should discuss the ladder of automation, the modes of control by human supervisors (in/on/out of the loop), and the legal status along the examples of autonomous cars, weapons, and medicine. (3) Based on the same use cases, the student should discuss technical/safety problems that hinder the full automation of processes.

Good starting points are:
Overview + Weapon Systems: I*CRC (2019). Autonomy, artificial intelligence and robotics: Technical aspects of human control, 1-28, https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control*
Medicine+ladder of automation (mainly p51+52): *Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med* ***25,*** *44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7*
Cars+ladder of automation: *Shladaver, S. E. (2016). The truth about "self-driving" cars. Scientific American, 314(6), 52-57 .https://www.scientificamerican.com/article/the-truth-about-ldquo-self-driving-rdquo-cars/*

Optional for conceptual insight:
Conceptual (First few pages): *Smithers, T. (1997). Autonomy in robots and other agents. Brain and cognition, 34(1), 88-106.*
Ethical Perspective (Section 2.7): *Müller, Vincent C., "Ethics of Artificial Intelligence and Robotics", The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/win2020/entries/ethics-ai/>.*
Human Control: *Tessier, C. (2017). Robots autonomy: Some technical issues. In Autonomy and Artificial Intelligence: A Threat or Savior? (pp. 179-194). Springer, Cham.*

Topic 6: Strategic classification [MSc] (Gunnar)

For a typical ML problem, the task definition starts with some input X and ends with a prediction Y. Of course, as ML systems learn from experience (X, Y) they are influenced by the outside world. This dependence is commonly taken into account, e.g. when analyzing the bias of the system.

Of course, ML systems that guide action, themselves influence the world. The linear process becomes a loop. A natural, understudied question, comes up: Taking the emerging dynamics into account, how should ML systems be designed?

A good starting point is:
Hardt, Moritz, et al. "Strategic classification." Proceedings of the 2016 ACM conference on innovations in theoretical computer science. 2016.

Topic 7: Model Safety, Uncertainty & Robustness [MSc] (Susanne)

Machine learning models are state-of-the-art in many fields, but they can be very sensitive to small changes in the data, for example, due to noisy data, measurement errors or artificial manipulations. The latter include adversarial examples/attacks where very small changes to an input lead to massive changes in the prediction of the machine learning algorithm.

The student should (1) give a definition of the robustness and safety of Machine Learning algorithms, (2) explain sources of uncertainty as well as measurements of uncertainty and (3) provide strategies against adversarial attacks and guarantees for robustness.

Good starting points are:
Faria, José. (2018). Machine Learning Safety: An Overview. Safety-critical Systems Symposium 2018 (SSS'18).
Adversarial attacks:

> *Liu, Qiang & Li, Pan & Zhao, Wentao & Cai, Wei & Yu, Shui. (2018). A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. IEEE Access. 6. 12103-12117. 10.1109/ACCESS.2018.2805680.*
> *Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning*
> *https://arxiv.org/abs/1712.03141*

Uncertainty:

> *Kläs, Michael. (2018). Towards Identifying and Managing Sources of Uncertainty in AI and Machine Learning Models - An Overview.*
> *https://arxiv.org/abs/1811.11669*

Topic 8: Deep Fakes [MSc, BSc] (Florian)

In recent years, the artificial generation of new text, images and videos has advanced to a state where it is often virtually impossible for humans to distinguish between what is real and what is fake. This has manifested itself in "deep fakes", e.g. inauthentic videos that show people saying or doing something different or put them in different sceneries and in inauthentic text generation, where e.g. systems pretend to be human and interact with humans on internet forums. In this topic, we aim to provide insight wrt. the current state of technology in those fields and the science behind those techniques, but simultaneously also the impact of such powerful technology in the societal and political discourse.

> *Deep Fakes, fake news bots, …*
> *https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy*

Detection of Deep Fakes or Fake News
https://techcrunch.com/2019/06/10/to-detect-fake-news-this-ai-first-learned-to-write-it/?guccounter=1
A GPT3 Bot went undetected on reddit
https://www.kmeme.com/2020/10/gpt-3-bot-went-undetected-askreddit-for.html

Korshunov, DeepFakes: a New Threat to Face Recognition? Assessment and Detection
https://arxiv.org/abs/1812.08685

Topic 9: Differential Privacy and Model Re-identification [BSc, MSc] (Florian)

Personal data that can be used for training machine learning models is highly sought after, but at the same time needs to be closely guarded as it contains sensitive data, e.g. a patient's history in a medical domain. Therefore, such types of data are awarded special protection in our laws. This calls for special, privacy preserving methods when data can e.g. not be shared between hospitals and simultaneously requires that, should trained models be made publicly accessible, unique training points can not be re-identified. This topic covers both aspects, briefly introducing the relevant concept of one and providing a more in-depth discussion for the other.

Differential Privacy

Kobbi Nissim, et al. Differential Privacy: A Primer for a Non-technical Audience. February 14, 2018.
Dwork, The reusable Holdout:
https://science.sciencemag.org/content/349/6248/636.abstract
FacebookAI, Introducing Opacus
https://ai.facebook.com/blog/introducing-opacus-a-high-speed-library-for-training-pytorch-models-with-differential-privacy

Re-Identification

Yogarajan, Vithya et al. “A review of Automatic end-to-end De-Identification: Is High Accuracy the Only Metric?” Applied Artificial Intelligence 34 (2020): 251 – 269.
El Emam K, Jonker E, Arbuckle L, Malin B (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLOS ONE 6(12): e28071.
https://doi.org/10.1371/journal.pone.0028071

Topic 10: Algorithmic Fairness - Technical vs. Societal Solutions [BSc, MSc] (Florian)

Automated decision making systems interact with humans on a daily basis and encroach many aspects of their lives. This ranges from the simple ordering of google search results and which prices are shown to me in web-store to whether my loan application is granted or not. Those systems on the other hand often exhibit systemic biases, such that parts of the affected population, often minorities suffer from higher error rates or an overall worse experience. In order to build fair(er) systems, we require ways to measure fairness in such systems. In this topic, we will examine several such metrics, their motivating principles but also dangers arising from solely relying on such metrics.

Hardt, Equality of Opportunity in Supervised Learning https://arxiv.org/abs/1610.02413
Kilbertus, Causal perspective on fairness https://arxiv.org/abs/1706.02744
Tutorial: 21 fairness definitions and their politics:
https://www.youtube.com/watch?v=jIXIuYdnyyk
Fairness and Abstraction in Sociotechnical Systems
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265913
Fairwashing: the risk of rationalization
https://arxiv.org/abs/1901.09749