ML4520 Philosophy and Ethics of Machine Learning

Seminar, University of Tübingen, Summer 2026

Instructors

Sebastian Zezulka [sebastian.zezulka@uni-tuebingen.de]
Prof. Bob Williamson

Office hour: by appointment. Please reach out if you have any questions.

Course Description

Algorithmic systems are tools of governance when they are used to determine how social roles are assigned and how social goods are distributed. They influence the support unemployed receive; how health care resources are allocated among patients; and whether prisoners are released or detained before trial. As companies and public administrations entrust increasingly significant predictions and decisions to algorithmic systems, data scientists and machine learning practitioners are empowered to determine aspects of the structure of our societies. Attending their rise in power and prestige is a growing chorus of critics calling for fairness, justice, and democratic accountability in algorithmic decision-making. These developments are, in their technological aspect, unprecedented. At the same time, the questions they raise concerning the just distribution of social goods and roles belong to the most traditional areas of philosophical inquiry.

This course introduces students to (1) recent technical literature in algorithmic fairness, (2) empirical evaluations of allocation mechanisms, and (3) classic philosophical work on the just distribution of social roles and goods. We stress the continuity between contemporary problems in algorithmic fairness and similar difficulties that predate machine learning. Engaging with literature drawn from a wide range of disciplines will present significant challenges but, we hope, yield equally significant rewards.

Course Requirements

To earn 3 ECTS (grade/ungraded) you have to

1. formally register for the course and for a class presentation at the end of the first week (19.04.2026). Please attend the introductory session to receive all information;
2. prepare both required readings before each class and actively participate in the discussions;
3. post one question about the readings in the forum no later than the day before class. You may miss up to two posts during the term.
4. present one 8-minute summary of a reading during class. No slides, focus on the key aspect of the paper. Schedule an office hour with me no later than one week before your presentation;
5. write a 1,500-word essay at the end of the term on a topic of your choice, related to the course.
    a. By Friday, 03.07.2026, submit a one-page outline for your essay topic that includes your thesis statement, planned argument structure, and key literature. You'll receive feedback based on this outline.
    b. The final submission deadline for the essay is Friday, 25.09.2026. Submissions are only accepted via Moodle and as pdf files.

To earn 6/8/12 ECTS, you have to write an essay of 3/4/5 thousand words, respectively.

Grading is determined as follows:

- Submission of reading questions: 10 %
- Presentation: 40 %
- Final essay: 50 %

Missing class and late assignments

We recognize that occasional problems associated with illness, family emergencies, job interviews, other professors, etc. will inevitably lead to legitimate conflicts over your time. If you expect that you will be unable to turn in an assignment on time, or must be absent from a class meeting, please notify us via email in advance, and we can agree on a reasonable accommodation.

Academic Integrity

Each student is responsible for being aware of the university policies on academic integrity, including the policies on cheating and plagiarism. When writing your essay, you are **not allowed** to use AI support for more than planning and copy-editing your written assignment.

In case we have a reasonable doubt about this, we will invite you to a follow-up discussion with one of the instructors to determine whether you have written the seminar paper independently. The follow-up discussion is not graded.

Background readings

These optional resources are provided to help you explore the topic and literature.

- Vredenburgh, Kate. "Fairness." in Justin B. Bullock, and others (Eds), The Oxford Handbook of AI Governance. 2022.
- Binns, Reuben. "Fairness in machine learning: Lessons from political philosophy." *Conference on Fairness, Accountability and Transparency*. PMLR, 2018.
- Hellman, Deborah, "Algorithmic Fairness", The Stanford Encyclopedia of Philosophy. Fall 2025 Edition.
  [https://plato.stanford.edu/entries/algorithmic-fairness/]
- Whittlestone, Jess, et al. “The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions.” *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.
- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2023.
  [https://fairmlbook.org/]

Seminar Plan and Reading List

April 13 – Introduction

No required readings.
Please register for the seminar and a presentation by April 19, 2026.

April 20 – The risk of long-term unemployment

- Sam Corbett-Davies, Sharad Goel. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” **(Version 2), 2018**.
  [arXiv:1808.00023v2]
- Achterhold, Eva, et al. "Fairness in Algorithmic Profiling: The AMAS Case." *Minds and Machines* 35.1, 2025.

April 27 – Health care resources

- Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366.6464, 2019.

- Tal, Eran. "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023.

May 04 – Criminal justice

- Chouldechova, Alexandra. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”. *Big Data* 5.2, 2017.
- Ludwig, Jens, and Sendhil Mullainathan. "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System." *Journal of Economic Perspectives* 35.4, 2021.

May 11 – School dropout

- Perdomo, Juan Carlos, et al. "Difficult lessons on social prediction from Wisconsin public schools." *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2025.
- Barabas, Chelsea, et al. "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment." *Conference on fairness, accountability and transparency*. PMLR, 2018.

May 18 – Legitimacy and Justice

- Kuppler, Matthias, et al. "From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making." *Frontiers in sociology* 7. 2022.
- Wang, Angelina, et al. "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy." *ACM Journal on Responsible Computing* 1.1, 2024.

May 25 – Whit monday

No class.

June 01 – Rawls: Justice as Fairness

- Rawls, John. "Justice as fairness." *The Philosophical Review,* 67.2,1958.
- Arrow, Kenneth J. "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice." *The Journal of Philosophy.* 1973.

June 08 – Against Rawls: Libertarians and Socialists

- Nozick, *Robert. Anarchy, State, and Utopia*. Chapter 7: Distributive Justice. 1974.

- Cohen, Gerald. A. “The Pareto argument for inequality.” *Social Philosophy and Policy* 12(1), 1995.

June 15 – Against Rawls: Theories of Race and Gender

- Mills, Charles. “Ideal Theory as Ideology.” *Hypatia*. 2005.
- Okin, Susan Moller. *Justice, Gender, and the Family.* Chapter 5: Justice as Fairness: For Whom? 1989.

June 22 – Justice beyond Distribution

- Young, Iris Marion. *Justice and the Politics of Difference.* Chapter 1. 1990.
- Kasirzadeh, Atoosa. "Algorithmic fairness and structural injustice: Insights from feminist political philosophy." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022.

June 29 – Local Justice

- Elster, Jon. "Local justice: how institutions allocate scarce goods and necessary burdens." *European Economic Review* 35.2-3. 1991.
- Dwork, Cynthia, et al. “Fairness through awareness.” Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 2012.

July 06 – Non-discrimination and the law

- Weerts, Hilde, et al. "Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023.
- Hu, Lily, and Issa Kohler-Hausmann. "What's sex got to do with machine learning?." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020.

July 13 – Can algorithms promote justice after all?

- Kleinberg, Jon, et al. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics.* 133(1). 2018.
- Vredenburgh, Kate. "AI and Bureaucratic Discretion." *Inquiry* 68.4. 2025.

July 20 – Troubles with allocating social goods

- Shirali, Ali, Rediet Abebe, and Moritz Hardt. "Allocation Requires Prediction Only if Inequality Is Low." International Conference on Machine Learning. PMLR, 2024.

- Zezulka, Sebastian and Konstantin Genin. Predictions, Performativity, and Potential Outcomes: Communicative Rationality in Prediction-Allocation Problems. EAMMO’25, 2025.