Seminar: Ethics and Philosophy of Machine Learning

Instructors
Thomas Grote [thomas.grote@uni-tuebingen.de]
Konstantin Genin [konstantin.genin@uni-tuebingen.de]

Meeting Time
Thursdays, 16-18h
https://zoom.us/j/3709690607

Course Description
Algorithmic systems increasingly determine how desirable and undesirable social roles are assigned and how desirable goods are distributed and withheld. They determine which job applicants receive consideration; which prisoners are released before trial; and which loan applicants receive bank credit. As algorithmic systems are entrusted with decisions of increasing significance, machine learners and data scientists are increasingly empowered to determine the fundamental structure of society. Attending their rise in power and prestige is a growing chorus of critics calling for fairness and democratic accountability in algorithmic decision-making. These developments are, in their technological aspect, entirely unprecedented. On the other hand, the question of how social goods ought to be distributed and how social roles ought to be assigned is one of the most traditional areas of philosophical inquiry. In this course, we introduce the student to both (1) cutting-edge technical literature in algorithmic fairness and (2) classic philosophical work on just distribution of social roles and goods. We stress the continuity of contemporary problems in algorithmic fairness to similar difficulties throughout history, including twentieth-century psychometric testing; tenth-century examinations for the promotion of scholar-bureaucrats in Song dynasty China; and Plato's ancient Greek blueprint for the just city-state. The combination of literature from a wide range of disciplines and time periods will both present significant challenges and, hopefully, yield significant rewards.

Course Requirements
Class will meet on Zoom, from 16-18h CET, every Thursday between April 19th and July 30th, with the exception of holidays on May 26th and June 9th and 16th, for a total of 12 class meetings. There will be required readings for every meeting. Everyone should make an effort to read all the material. We will provide you with the materials, but if there is some difficulty please make an effort to find the material yourself. The majority of class time will be devoted to student presentations of the assigned readings. Every student must present exactly one reading. Presenters may take on the extra responsibility of background reading for the material they are presenting. Readings will be assigned at the first meeting accounting to some degree for student preference. Presenters should make an effort to present the material in the readings as charitably, clearly and succinctly as possible. The presentations should last 15-20 minutes, allowing for 10-15 minutes of discussion. To allow for as much discussion as possible, presenters should make an effort to come in at the shorter end of the 15-20 minute range.

Except for a handful of exceptions, there will be two student presentations per class. The instructors will make themselves available beforehand to discuss the material for the presentation. There may be up to 30 minutes devoted to lectures by either Thomas or Konstantin.

There will also be a 1,500 word essay due on **September 9th**. The subject matter is flexible and intended to answer to individual interest, but students must submit a 1-page proposal for approval by **July 15th**.

Grading is determined as follows:

Class participation: 10%
Presentation: 45%
Final essay: 45%

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. Otherwise, your grade will be penalized.

Academic Integrity
It is the responsibility of each student to be aware of the university policies on academic integrity, including the policies on cheating and plagiarism.

Reading List

Session 1 (April 21): **Introduction**

Session 2 (April 28): **ProPublica Sets the Agenda, Northpointe Responds**

- ProPublica (2016), *Machine Bias.* [article] [method] [risk assessment questionnaire]
- NorthPointe Inc. (2016), *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. [article]

Session 3 (May 5): **You Can't (Ever) Get What You Want: Fairness Trade-Offs**

- Kleinberg et al. (2016), *Inherent Trade-Offs in the Fair Determination of Risk Scores.* [article]
- Ra [article]

Session 4 (May 12): **Footnotes to Plato, Or: How to Cultivate Philosopher-Kings**

- Plato (circa 375 BC), *Republic.* [Book II] [Book III]

Session 5 (May 19): **Pick the Very Best One: Trouble in the Meritocracy**

- Chaffee (1985), *The Thorny Gates of Learning in Sung China: A Social History of Examinations.* [Chapters 2-3, see **ILIAS**]
- Popper (1945), *The Open Society and Its Enemies.* [Chapter 7. The Principle of Leadership]
- C.L.R. James (1956), "Every Cook Can Govern". [article]

Session 6 (June 2): **Rawls: Justice as Fairness**

- Rawls (1958), *Justice as Fairness*. [article (also on **ILIAS**)]
- Wenar (2021), *John Rawls*. [SEP]

Session 7 (June 23) **Against Rawls: Libertarians and Socialists**

- Nozick (1974), *Anarchy, State, and Utopia.* [Excerpts, Chapter 7 (see **ILIAS**)]
- Cohen (1995), "The Pareto Argument for Inequality". [article (also on **ILIAS**)]

Session 8 (June 30) **Against Rawls: Theories of Race and Gender**

- Mills (2005) "Ideal Theory as Ideology" [article (also on **ILIAS**)]
- Okin (1991), *Justice, Gender, and the Family.* [Chapter 5: Justice as Fairness: For Whom? (see **ILIAS**)]
- Crenshaw (1991) "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color" [article (also on **ILIAS**)]

Session 9 (July 7) **Algorithmic Fairness Again: Groups, Small Groups and Individuals**

- Kearns et al. (2019) "An Empirical Argument for Rich Subgroup Fairness in Machine Learning" [article]
- Dwork et al. (2012) "Fairness Through Awareness" [article]

Session 10 (July 14) **Can Algorithms Promote Justice After All?**

- Kleinberg et al. (2018) "Human Decisions and Machine Predictions" [article]
- Ludwig and Mullainathan (2021) "Fragile Algorithms and Fallible Decision-Makers" [article]

Session 11 (July 21) **Algorithmic Fairness: Causal Fairness**

- Kilbertus et al. (2017) “Avoiding Discrimination through Causal Reasoning” [article]
- Hu and Kohler-Hausman (2020) “What's Sex Got To Do With Fair Machine Learning?” [article]

Session 12 (July 28) **Algorithmic Fairness: The View from Psychometrics**

- Borsboom et al (2008) “Measurement invariance versus selection invariance: is fair selection possible?” [article]
- Hutchinson and Mitchell (2018) “50 Years of Test (Un-)Fairness: Lessons for Machine Learning” [article]