Introduction to Fairness, Accountability, and Transparency in Machine Learning

1. Course Description

The purpose of this course is to equip you with the technical and conceptual tools to be able to engage deeply and critically with machine learning research and practice. The course strives to integrate historical and cultural context with contemporary methods in machine learning and scholarship on Fairness, Accountability, and Transparency (FAT*). The basic strategy is to ground the course in issues of public concern related to machine learning and build practical knowledge through engagement with the technical workings of algorithms related to those concerns.

The core goal of each class will be to ask, in addition to the mathematical or statistical assumptions being made by a given machine learning model, what are the social and psychological assumptions associated with the use ofthat model in particular situated applications? And what are the varied consequences of those assumptions in those particular social contexts?

Course objectives:

- Gaining perspective of machine learning technologies and practice
- Situating technology and practice in social relations of power, including patterns of discrimination and power asymmetry.
- Gaining an understanding of the institutional and historical contexts associated with key machine learning algorithms.
- Building or identifying tools for reflective practice and radical interventions in machine learning.

2. Assignments

The primary assignment for this course is an essay on a topic of your choosing. Essay topics can range from describing a conceptual or technical contribution of your own to presenting a thorough and informed critique of an existing conceptual or technical contribution. You may optionally include a coding component according to your personal learning objectives. The essay requires that you, first describe the technical or conceptual contribution in detail and then critically assess it as socio-technical phenomenon, situating it within appropriate literature.

3. Activities and Topics

1. Introduction
2. Introductory Machine Learning – Targeted Advertising and Consumer Scoring
3. Neural Networks – Surveillance and Capture
4. Dependent Data – The Global Financial Crash
5. Deep Learning – The Ethics of Inferential AI
6. Causal Inference – The Scope of Digital Public Policy

7. PageRank – Actor Network Theory
8. Probabilistic Programming – Reconfiguration & Collective Liberation

4. Readings (to be taken from this list)

Agre, P. E. (1994). Surveillance and capture: Two models of privacy. The Information Society, 10(2), 101–127. https://doi.org/10.1080/01972243.1994.9960162

Alexander, M. (2010) The new Jim Crow: mass incarceration in the age of colorblindness. New York: New Press.

Ames, M. G. (2015). Charismatic Technology. Aarhus Series on Human Centered Computing, 1(1), 12.

Benjamin, R. (2019a) Race after technology: abolitionist tools for the new Jim code. Medford, MA: Polity.

Bowker, G. C. and Star, S. L. (1999) Sorting things out: classification and its consequences. Cambridge, Mass: MIT Press.

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Breyman, S. et al. (2017) 'STS and Social Movements: Pasts and Futures', in Felt, U. et al. (eds) The handbook of science and technology studies. Fourth edition. Cambridge, Massachusetts: The MIT Press.

Broussard, M. (2018) Artificial unintelligence: how computers misunderstand the world. Cambridge, Massachusetts: The MIT Press.

DiSalvo, C. (2012). *Adversarial design*. Cambridge, Mass: MIT Press.

Bardzell, S., Bardzell, J., Forlizzi, J., Zimmerman, J., & Antanitis, J. (2012). Critical Design and Critical Theory: The Challenge of Designing for Provocation. Proceedings of the Designing Interactive Systems Conference, 288–297. https://doi.org/10.1145/2317956.2318001

Domingos, P. (2015) The master algorithm: how the quest for the ultimate learning machine will remake our world. [London]: Allen Lane.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, *8*(1), 19–30. https://doi.org/10.1007/s00779-003-0253-8

Eubanks, V. (2017) Automating inequality: how high-tech tools profile, police, and punish the poor. New York, NY: St. Martin's Press.

Friedman, B. and Hendry, D. (2019) Value sensitive design: shaping technology with moral imagination. Cambridge, Mass: MIT Press.

Giddens, A. (1986) Sociology: a brief but critical introduction. 2nd ed. Basingstoke: Macmillan Education.

Gray, M. L. and Suri, S. (2019) Ghost work: how to stop Silicon Valley from building a new global underclass. International edition. Boston: Houghton Mifflin Harcourt.

Haraway, Donna (1988) 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', Feminist Studies, 14(3), pp. 575–599.

Hicks, M. (2018) Programmed inequality: how Britain discarded women technologists and lost its edge in computing. Cambridge, MA: MIT Press.

Irani, L. (2019) Chasing innovation: making entrepreneurial citizens in modern India. Princeton, NJ: Princeton University Press.

Jasanoff, S. (2004) 'The idiom of co-production', in States of knowledge: the co-production of science and social order. London: Routledge.

Latour, B. (2005) Reassembling the social: an introduction to actor-network-theory. Oxford: Oxford University

Mittelstadt, B. D. et al. (2016) 'The ethics of algorithms: Mapping the debate', Big Data & Society, 3(2). doi: 10.1177/2053951716679679.

Noble, S. U. (2018a) Algorithms of oppression: how search engines reinforce racism. New York: New York University Press.

O'Neil, C. (2017) Weapons of math destruction: how big data increases inequality and threatens democracy. United Kingdom: Penguin Books.

Pasquale, F. (2015) The black box society: the secret algorithms that control money and information. Cambridge, Massachusetts: Harvard University Press. Available at: http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=3301535.

Rosenberg, A. and Curtain, T. (2018) Philosophy of Social Science. 5th ed. Boulder: Routledge.

Rosenblat, A. (2018) Uberland: how algorithms are rewriting the rules of work. Oakland, California: University of California Press.

Suchman, L. (2008) 'Feminist STS and the Sciences of the Artificial', in The handbook of science and technology studies. 3rd ed. Cambridge, Mass: MIT Press.

Wajcman, J. (2004) TechnoFeminism. Cambridge: Polity.

Zuboff, S. (2019) The age of surveillance capitalism: the fight for the future at the new frontier of power. London: Profile Books.