Module Proposal

Accountable AI: Making data lives equitable

Module description

This course will provide students with interdisciplinary approaches to understand the social and ethical effects of the ubiquitous deployment of AI technology and Machine Learning. This includes (but is not limited to) biases in data processing, surveillance, privacy, misrepresentation, and recourse. Machine Learning and AI systems seem to generate "automated decisions", but people are central at every step of the lifecycle: defining tasks, generating datasets, preprocessing and labelling, which models to use, what to optimize, and so on. We will engage with the socio-technical nature of these issues in order to think through the needs for accountable and human-centered AI. We also critically investigate the regimes of knowledge in computer science when dealing with AI accountability.

We will start with contemporary and intersectional foundations in ethics, and then move to the current and growing research literature on ethics in AI, before considering specific AI tasks, data sets and methodologies through the lens of the ethical considerations identified. The weekly discussions are grounded in the implementation of AI in "data life": the everyday interactions of citizens with predictive AI technology and big data. In doing so, students will become acquainted with different perspectives on AI accountability. This incudes ethical orientations towards accountability, legal frameworks for AI implementation, and computational methods for fairness and transparency.

Students will learn to formulate an interdisciplinary perspective on AI through multiple theoretical frameworks grounded in technical and humanistic perspectives, and to construct and explain logical arguments about AI in relation to significant issues in studies of technology, data ethics, literature, culture, and society. The course will feature guest lecturers from different disciplinary backgrounds to ensure a diversity of standpoints, voices and concerns.

Assessment

The module is assessed based on a weekly blog post (300w, 30%) in which you detail your daily experiences with AI technology from a perspective of ethics and accountability.

For the final assignment of 4,000 words (70%), you have two options.

1. A paper analyzing some particular AI system or data set in terms of the concepts developed, and looking forward to how ethical best practices could be developed for that task or dataset. In the paper, you will ask questions such as "what went wrong?" and "who was harmed?", but also "who benefitted?" and "how could we mitigate his kind of harm in the future?" from ethical, legal, and technical perspectives.
2. An essay dealing with the future of accountable AI. Using academic sources, construct an argument about the future of AI from an ethical point of view. Using a particular example of AI, either defend or challenge this use of AI, and offer arguments regarding future developments of this particular use case. Think specifically about the ways in which the system can be made more accountable from the perspective of individual users, and what kinds of dangers or challenges lie ahead.

Module Materials and Student Participation

Readings, lecture slides, notes, will be available on the module's KEATS page. Each session will have several dedicated readings, will we will approach using a 'divide and share' tactic: everyone reads something else, with the same questions in mind. You will generally be asked to pick readings that offer a different perspectives to your own. During seminars, you will share your findings with the rest of class based on the questions for each week (see below).

Week 1

Introduction

We will discuss why we are here, and what we hope to accomplish in this class. Each week will be grounded in specific examples of AI in our "data lives", which we will discuss during the seminars. Before this first seminar, **read two of the calls to action below**, and write down what is at stake – socially and ethically – in the widespread use of AI. Bring your notes to the seminar.

Readings (pick 2)

- Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538 (7625), 311.
- O'Neil, C. (Sep 8, 2020). Mutant Algorithms Are Coming for Your Education. Bloomberg.
- Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 591-598). Berlin, Germany: Association for Computational Linguistics.
- Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. Proceedings of the National Academy of Sciences, 113 (48), 13538-13540.

- Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), Media technologies: Essays on communication, materiality, and society (pp. 167-194). MIT Press.
- Sourour, B. (Nov 13, 2016). The code I'm still ashamed of. Medium.com.
- Knight, W. (Nov 19, 2019). The Apple Card Didn't 'See' Gender—and That's the Problem. Wired.com
- Kayser-Bril , N. (Feb 6, 2020). The algorithm police is coming. Will it have teeth? Algorithm Watch.

Week 2

Thinking about ethics

Today, we will discuss general ethical concerns from a philosophical perspective. Each of you will **read two items from the reading list below,** at least one of which comes from an author whose perspective varies significantly from your own life experience. Be prepared to discuss the following questions during the seminar:

- What is the main thesis of the reading?
- Which definition of ethics and/or accountability is used?
- How does this definition contrast with others?
- How would you relate this reading to the issue of accountability in AI?

Readings (pick 2)

- Bartky, S. L. (2002). "Sympathy and solidarity" and other essays (Vol. 32). Rowman & Littlefield.
- Singer, P. (2011). *Practical Ethics*. Cambridge University Press.
- Bryson, J. J. (2015). Artificial intelligence and pro-social behaviour. In C. Misselhorn (Ed.), Collective agency and cooperation in natural and artificial systems: Explanation, implementation and simulation (pp. 281-306). Cham: Springer International Publishing.
- Butler, J. (2005). Giving an account of oneself. Oxford University Press.
- hics: A liberative approach. Fortress Press.
- Edgar, S. L. (2003). Morality and machines: Perspectives on computer ethics. Jones & Bartlett Learning.
- Dubber, M. D., Pasquale, F., & Das, S. (2020). *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press.
- Quinn, M. J. (2013). *Ethics for the Information Age*. Boston: Pearson.
- Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical Considerations in Artificial Intelligence Courses. *AI Magazine* 38: 2.
- Fieser, J., & Dowden, B. (Eds.). (2016). Internet encyclopedia of philosophy: Entries on Ethics
- Liamputtong, P. (2006). Researching the vulnerable: A guide to sensitive research methods. Sage.
- Metcalf, J., Keller, E. F., & boyd, d. (2016). Perspectives on big data, ethics, and society. *Council for Big Data, Ethics, and Society.*

Week 3

Accountable AI
Today, we will look at common ethical issues that accompany AI decision-making—particularly, the question who can be held accountable in AI deployment. This question is a complex one, without a one size fits all solution. **Pick two readings** from the list below, and summarize the following:

- What are the main problems discussed?
- What are the main solutions offered?
- Do you think these solutions are sufficient? If not, why?

Readings (pick 2)

- Umoja Noble, Safiya (2018). *Algorithms of oppression*. New York: New York University Press. Chapter 1.
- Horning, R. (2017). *Sick of Myself.* Real Life Mag. https://reallifemag.com/sick-of-myself/
- Diakopoulos, N. (2015). Algorithmic Accountability : Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415.
- Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 146144481667664.
- O'Neil, Cathy (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. NY: Crown Publishing Group. Chapter 1.
- Benjamin, Ruha (2019). *Race after Technology*. Cambridge: Polity. Chapter 1.
- Wachter-Boettcher, Sara (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. New York & London: W. W. Norton & Company. Chapter 1.
- Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy and Technology*, *33*(4), 685–703.
- Broussard, M. (2018). *Artificial Unintelligence*. Cambridge & Malden: MIT Press. Chapter 1.
- Brock Jr., A. (2020). *Distributed Blackness*. New York: New York University Press. Chapter 1.

Week 4
Legal systems and AI
Today, we will discuss the current legal frameworks that exist in the UK and EU in order to handle discriminatory AI systems, such as the GDPR. However, "one size fits all" regulation can be problematic and cause backlash. Choose **one of the three legal regulations** surrounding the use of big data, and pick **one reading on legal issues** surrounding AI. We will ask ourselves the following questions:

- Which kinds of big data usage and AI decision-making are targeted by law?
- Which problematic aspects of big data and AI are not included?
- How are the legal issues in the literature related to the ethical issues you explored last week?
- To which extent is legal regulation fit to deal with AI decision-making? Why?

Legal regulations (pick 1)

- UK Data Protection Act
- EU GDPR
- China's Cybersecurity Law

Readings (pick 1)

- Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A.-C., Levin, F., & Rondepierre, B. (2018). For a meaningful Artificial Intelligence: Towards a French and European strategy, 152.
- Hildebrandt, M., & Koops, B.-J. (2010). The Challenges of Ambient Law and Legal Protection in the Profiling Era. *The Modern Law Review Limited*, *73*(3), 428–460.
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., … Wood, A. (2016). Accountability of AI Under the Law: The Role of Explanation.
- Aletras, N., Tsarapatsanis, D., Preoțiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European court of human rights: A natural language processing perspective. *PeerJ Computer Science*, *2016*(10), 1–19. https://doi.org/10.7717/peerj-cs.93
- Alarie, B. (2016). The Path of the Law: Toward Legal Singularity. *SSRN Electronic Journal*, *10*(1897), 1–11. https://doi.org/10.2139/ssrn.2767835
- Liu, J. Z., & Li, X. (2019). Legal Techniques for Rationalizing Biased Judicial Decisions: Evidence from Experiments with Real Judges. *Journal of Empirical Legal Studies*, *16*(3), 630–670. https://doi.org/10.1111/jels.12229
- Rice, D. (2019). Understanding Legal Meaning through Word Embeddings. *SSRN Electronic Journal*, 0–29. https://doi.org/10.2139/ssrn.3455747
- Katz, D. M. (100AD). Quantitative Legal Prediction, *1*(July 2011). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2187752

Week 5

Digital Discrimination

We will look into different examples of discriminatory AI. Many current technologies do not seem to be tested with non-white people, such as face tracking technologies, fitness trackers with optical sensors, and commercial gender recognition apps. Each of you will pick **two readings from the list below** and ask yourself:

- What went wrong?
- Who was harmed?
- Who benefitted?
- What (if anything) is proposed to mitigate this harm in the future?
- How would legal frameworks help to deal with this problem? Why (not)?

Readings (pick 2)

- Angwin, J., & Larson, J. (Dec 30, 2016). Bias in criminal risk scores is mathematically inevitable, researchers say. ProPublica.
- Benjamin, R. (2019). 'Assessing risk, automating racism', *Science*, Vol. 366, Issue 6464, pp. 421-422.

- Hao, K. (2019). 'Facebook's ad-serving algorithm discriminates by gender and race', *MIT Technology Review*.
- Kirchner, Lauren. 2015. When Discrimination Is Baked Into Algorithms. *The Atlantic.*
- Clark, J. (Jun 23, 2016). Artificial intelligence has a `sea of dudes' problem. Bloomberg Technology.
- Crawford, K. (Apr 1, 2013). The hidden biases in big data. Harvard Business Review.
- Emspak, J. (Dec 29, 2016). How a machine learns prejudice: Artificial intelligence picks up bias from human creators--not from hard, cold logic. Scientific American.
- Guynn, J. (Jun 10, 2016). `Three black teenagers' Google search sparks outrage. USA Today.
- Hardt, M. (Sep 26, 2014). How big data is unfair: Understanding sources of unfairness in data driven decision making. Medium.
- Jacob. (May 8, 2016). Deep learning racial bias: The avenue Q theory of ubiquitous racism. Medium.
- Larson, J., Angwin, J., & Parris Jr., T. (Oct 19, 2016). Breaking the black box: How machines learn to be racist. ProPublica.
- Morrison, L. (Jan 9, 2017). Speech analysis could now land you a promotion. BBC capital.
- Sweeney, L. (May 1, 2013). Discrimination in online ad delivery. Communications of the ACM, 56 (5), 44-54.

Week 6

Language Bias and Word Embeddings

Today, we will discuss language biases and their integration into word embedding models. In these models, human biases are taken up by computational systems. **Pick two readings from the list below** and, like last week, ask the following questions:

- What went wrong?
- Who was harmed?
- Who benefitted?
- What (if anything) is proposed to mitigate this harm in the future?
- How would legal frameworks help to deal with this problem? Why (not)?

Readings (pick 2)

- Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. CoRR, abs/1607.06520.
- Caliskan-Islam, A., Bryson, J., & Narayanan, A. (2016). A story of discrimination and unfairness. (Talk presented at HotPETS 2016)
    - Video of presentation
- Daumé III, H. (2016). Language bias and black sheep.
- Herbelot, A., Redecker, E. von, & Müller, J. (2012, April). Distributional techniques for philosophical enquiry. In Proceedings of the 6th workshop on language technology for cultural heritage, social sciences, and humanities (pp. 45-54). Avignon, France: Association for Computational Linguistics.

- Schmidt, B. (2015). Rejecting the gender binary: A vector-space operation.
- Fessler, Leah. (Feb 22, 2017). SIRI, DEFINE PATRIARCHY: We tested bots like Siri and Alexa to see who would stand up to sexual harassment. Quartz.
- Fung, P. (Dec 3, 2015). Can robots slay sexism? World Economic Forum.
- Mott, N. (Jun 8, 2016). Why you should think twice before spilling your guts to a chatbot. Passcode.
- Paolino, J. (Jan 4, 2017). Google home vs Alexa: Two simple user experience design gestures that delighted a female user. Medium.
- Seaman Cook, J. (Apr 8, 2016). From Siri to sexbots: Female AI reinforces a toxic desire for passive, agreeable and easily dominated women. Salon.
- Twitter. (Apr 7, 2016). Automation rules and best practices.

Week 7

Surveillance and Policing

Today we will look at the ways in which AI is used in policing and surveillance contexts. Some hold that "if other things equal, shoe size is a useful predictor of recidivism, then it can be included as a predictor" (Berk et al 2013). **Pick two of the examples from the list below** about AI surveillance systems across the world (UK, US, China), and answer the following questions we will discuss in class:

- Who is being policed?
- Who does the policing benefit?
- What legal recourse do the policed have?
- Which factors should (not) be used in these predictive systems? Why?

Readings (pick 2)

- https://urbanomnibus.net/2018/06/non-exhaustive-taxonomy-tools-data-driven-policing/
- Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. Criminology & Pub. Pol'y, 12, 513.
- Ferguson, A. G. (2017). *The rise of big data policing*. New York University Press. Chapter 1.
- Shepherd, S. M., & Willis-Esqueda, C. (2018). Indigenous perspectives on violence risk assessment: A thematic analysis. *Punishment and Society*, *20*(5), 599–627.
- Gillingham, P. (2016). Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the "Black Box" of Machine Learning. *British Journal of Social Work*, *46*(4), 1044–1058.
- Sausdal, D. (2019). Policing at a distance and that human thing: An appreciative critique of police surveillance. *Focaal*, *2019*(85), 51–64.
- Weitzer, R., Brewer, J. D., Guelke, A., Hume, I., Moxon-Browne, E., & Wilford, R. (1989). The Police, Public Order, and the State: Policing in Great Britain, Northern Ireland, the Irish Republic, the USA, Israel, South Africa, and China. *Contemporary Sociology* (Vol. 18).
- Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. *Ssrn*, *222*(2015), 59–71.

Week 8

AI and Privacy

Today, we will discuss the issues of privacy surrounding AI implementation. From the following readings, **pick two (one from a technical perspective, and one from a social/legal perspective**). Ask yourself the following:

- How is privacy defined; which ethical frameworks is the author using?
- What purpose does privacy serve?
- Is there a tension between privacy and other values? If so, which?
- What unique concerns are there in NLP and privacy?

Readings (pick 2)

- Abadi, M., Chu, A., Goodfellow, I., Brendan McMahan, H., Mironov, I., Talwar, K., et al. (2016). Deep Learning with Differential Privacy. ArXiv e-prints.
- Amazon.com. 2017. Memorandum of Law in Support of Amazon's Motion to Quash Search Warrant
- Hawlitschek, F., Teubner, T., & Weinhardt, C. (2016). Privacy in the Sharing Economy. *Die Unternehmung*, *70*(1), 26–44.
- Such, J. M. (2017). Privacy and autonomous systems. *IJCAI International Joint Conference on Artificial Intelligence*, 4761–4767.
- Brant, T. (Dec 27, 2016). Amazon Alexa data wanted in murder investigation. PC Mag.
- Friedman, B., Kahn Jr, P. H., Hagman, J., Severson, R. L., & Gill, B. (2006). The watcher and the watched: Social judgments about privacy in a public place. Human-Computer Interaction, 21(2), 235-272.
- Golbeck, J., & Mauriello, M. L. (2016). User perception of facebook app data access: A comparison of methods and privacy concerns. Future Internet, 8(2), 9.
- Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information". Communications of the ACM, 53 (6), 24-26.
- Solove, D. J. (2007). 'I've got nothing to hide' and other misunderstandings of privacy. San Diego Law Review, 44 (4), 745-772.
- Mendelson, A. (2019). Security and Privacy in the Age of Big Data and Machine Learning. *Computer*, *52*(12), 65–70.

Week 9

Towards accountable AI

Today, we will have a look at different ways in which AI can be used in more responsible manner, in order to make our data lives more equitable. A human-centred, ethical, and accountable approach to AI means more than putting a "human in the loop". It touches upon aspects such as explainability, transparency, inclusivity—values which are not always compatible. Read at least **two papers from the reading list below**, and answer the following questions about these readings:

- What kinds of values are central here?
- How could this approach be applied to the discriminatory AI we have discussed?
- Are there tensions between these values and legal problems we have discussed?

Readings (pick 2)

- Bartlett, M. (2019, April 5). Solving the AI Accountability Gap. *Towards Data Science.*
- Busuioc, M. (2020). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, 1–12.
- Wallach, H. (Dec 19, 2014). Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. Medium.
- Wattenberg, M., Viégas, F., & Hardt, M. (Oct 7, 2016). Attacking discrimination with smarter machine learning.
- Ratto, M. (2011). Critical making: Conceptual and material studies in technology and social life. The Information Society, *27* (4), 252-260.
- Borning, A., & Muller, M. (2012). Next steps for Value Sensitive Design. *Conference on Human Factors in Computing Systems - Proceedings*, 1125–1134.
- Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artifcial intelligence. AI Magainze.
- Shilton, K., & Anderson, S. (2016). Blended, not bossy: Ethics roles, responsibilities and expertise in design. Interacting with Computers.
- Shilton, K., & Sayles, S. (2016). "We aren't all going to be on the same page about ethics": Ethical practices and challenges in research on digital and social media. In 2016 49th Hawaii international conference on system sciences (HICSS) (pp. 1909-1918).
- Leidner, J. L., & Plachouras, V. (2017). Ethical by Design: Ethics Best Practices for Natural Language Processing. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing.* Valencia.

Week 10

Next steps

In this final session, we will consider several proposals for ethical research in AI. Read one of the texts below, and ask yourself:

- What is missing in this proposal? Why?
- Which points should be reconsidered? Why?
- Should the academic AI community adopt a code of ethics of this general sort? Why or why not?

We will have a final discussion on the future of ethical AI and ML.

Readings (pick 1)

- ACM Ethics Task Force. (2018). ACM Code of Ethics and Professional Conduct.
- The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. (2016). Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems (AI/AS).
- Etlinger, S., & Groopman, J. (2015). The trust imperative: A framework for ethical data use.