PhD course on AI and Ethics

Proposal v. 07.01.2021

Course structure

- Intensive 2 weeks course
- 7 modules
- Hours per module 4 x 45mins

Learning aims

- For non-computer scientists: identify the way in which algorithms malfunction: do they malfunction for individuals, groups, in a given context, on a given task. This helps algorithm designers address the problem.
- For computer scientists: understand the implications of AI for society and organizations, learn the current debates about the design and use of AI, develop a critical understanding on AI.

Module 1: Introduction to Artificial Intelligence

Lecturer	AssProf Marija Slavkovik, UiB
Content	The module focuses on knowledge representation, reasoning and machine learning which are the areas of AI involved in automating decision-making.
Learning Aims	Learn the foundations and state-of-the-art in AI; undertsand the research goals and research methods of AI; ensure a common level of skills and understanding among all participants to enable them to follow the rest of the modules.
Exercise and aim	Make a spam filter using pen and paper. Aim: Learn the how problems are solved with AI in contrast with how people do it, to understand the advantages and limitations od AI.
Reading List	Appendix of "Human Compatible" by Stuart Russel; Chapters 1, 2 & 3 An Introductory Guide for Social Scientists by George David Garson (1998)
Additional literature	David Poole and Alan Mackworth. 2017. Artificial Intelligence: Foundations of Computational Agents (2 ed.). Cambridge University Press, Cambridge, UK. http://artint.info/2e/html/ArtInt2e.html

Module 2: Introduction to Artificial Intelligence from an non technical perspective

Lecturer	AssProf Miria Grisot, UiO; *AssProf Taina Bucher, UiO*
Content	The module focuses on understanding the curremt debates on the role of AI in society and in organizations
Learning Aims	Learn the current debates on the politics of AI; learn the current debate on the implications of AI and a good AI society, relation between AI and democracy, issues of surveillance and segmentation.
Exercise and aim	Discussion
Reading List (suggested)	Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review. Jobin, A., Ienca, M. & Vayena, E. (2019). 'The global landscape of AI ethics guidelines'. Nature Machine Intelligence, 1(9), pp. 389-99. Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. Ethics and Information Technology, 20(1): 5-14. Striphas, T. (2015). Algorithmic culture. European Journal of Cultural Studies, 18(4–5): 395–412.
Additional literature	Bucher, T. (2018). If... Then: Algorithmic power and politics. Oxford University Press. Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. Big Data & Society, 3(2). Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. Government Information Quarterly, 33(3): 371–377. Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology, 30(1), 75-89.

Module 3: Accountability and Transparency

Lecturer	Part I and II: Guest lecturers Part III: Miria Grisot
Content	Accountability and transparency
Learning Aims	Learn the foundations and state-of-the-art in accountaibility and transparency of AI. Get an entry point to research in this area: learn how to learn more and how to engage with that research community. Learn about the debate on responsibility and accountability related to AI design and deployment. AI accountability and the changing nature of work and organizing.
Exercise and aim	
Reading List (suggested)	Maranke Wieringa. 2020. What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* ’20). Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833 Nicholas Diakopoulos. 2020. Transparency. In The Oxford Handbook of Ethics of AI, Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.11 Floridi, L. (2019). Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 1(6), 261-262. Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), pp. 1-12.
Additional literature	tbd

Module 4: Explainable AI

Lecturer	Guest lecturers
Content	Explainable AI
Learning Aims	Understanding explainability from a user perspective: what is the expalinability problem. Learn what the basic problems and approaches are to making algorithms explainable. Understand the difference between explainability and interpretability of algorithms. Learn to compare and evaluae different ML algorithms with respect to their explainability.
Exercise and aim	For computer science students: the AIX360 toolkit. For non computer science students: analysis of examples of algorithmic explanataions (need to be identified!). Aim: Hands on experience with creating explanations for algorithms in ML
Reading List	Miller (2019); Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence Volume 267, February 2019, Pages 1-38 Arya u. a. (2019), Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilovi´c, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., and Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. Gunning und Aha (2019). David Gunning and David Aha. 2019. DARPA’s Explainable Artificial Intelligence (XAI) Program. AI Magazine 40, 2 (Jun. 2019), 44–58. https://doi.org/10.1609/ aimag.v40i2.2850
Additional literature	AIX360 kit, Video: Explainability 360 Tutorial by Amit Dhurandhar, September 18, 2019.5

Module 5: Algorithmic Fairness and bias mitigation

Lecturer	Guest lecturers
Content	Fairness and bias
Learning Aims	Learn what the concept of fairness means with respect to algorithms. Learn to recognise the different definitions of fairness, their motivation, strenghts and weaknesses. Learn the basic methods for mitigating bias in algorithms and data (pre-processing, in-processing and post processing)
Exercise and aim	For computer science students: the AIF360 toolkit. For non computer science students: analysis of examples of algorithmic bias (need to be identified!) Aim: Hands on experience with analysing and mittigating bias for algorithms in ML
Reading List	Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. Commun. ACM 63, 5 (2020), 82–89. https: //doi.org/10.1145/3376898 Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. CoRR abs/1908.09635 (2019). arXiv:1908.09635 http://arxiv.org/abs/1908.09635 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. CoRR abs/1104.3913 (2011). arXiv:1104.3913 http://arxiv.org/abs/1104.3913
Additional literature	Arvind Narayanantutorial at FAT2018. Trusted AI and AI Fairness 360 Tutorial by Prasanna (new tutorial from IJCAI 2020 will be made available) Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. CoRR abs/1810.01943 (2018). arXiv:1810.01943 http://arxiv.org/abs/1810.01943

Module 6: Privacy

Lecturer	Guest Lecturers
Content	Privacy
Learning Aims	The role that privacy concerns play in artificial intelligence. In particular the students will be introduced to the basic principles and methods of ensuring differential privacy and data.
Exercise	TBD
Reading List	Chapter 1 Kearns, M.; and Roth, A. 2019. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press. Datatylsinet (2018) Datatylsinet. 2018. Artificial intelligence and privacy. https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.
Additional literature	Dwork, C. 2006. Differential Privacy. In Bugliesi, M.; Preneel, B.; Sassone, V.; and Wegener, I., eds., Automata,Languages and Programming, 1–12. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-35908-1.

Module 7: other topics

Reserved for discussing open research problems in AI ethics, challenges and possible approaches.