Responsible Machine Learning: Legal and Technical Aspects

Course Syllabus

version on 10/18/2023

Doaa Abu Elyounes and Nicolas Vayatis

Teaching assistant- Marie Garin

Date and time: Wednesdays 08.30-10.30

Location : Faculté de médecine - Site Cochin

Course description

This course will examine the mutual relationship between law, computer science and public policy. With the wide spread of artificial intelligence, the link between these domains is becoming more and more ambiguous.

On one hand, the growth in computing capacity combined with the growth in the amount of data collected has revealed the exceptional modeling and data processing capabilities of machine learning. Great advances have been made for a series of specialized tasks such as speech recognition, translation, decision support systems, fault detection, image classification, etc. opening the door to many applications in health, industry, social sciences and many other fields. On the other hand, the automation of certain tasks or decisions raises a number of significant issues, such as the exacerbation of structural inequalities whereby in domains like predictive justice, hiring and finance, algorithms are leading to decisions that benefit mainly those who are already in power and further disadvantage minorities and vulnerable groups. Although automation is often thought of as a "simple" substitution, it reinforces existing biases because it encodes certain norms into complex lines of code that "obscure" the true nature of the problem (Carr, 2017; Dekker, & Woods, 2002, Surveillance Capitalism Zuboff, 2019). In addition, recommendation systems also embody harmful effects such as political destabilization, mental health issues, misinformation and disinformation. The resource intensive nature of the technology has also an effect on the

the environment and ecosystems (Strubell et al, 2019). Recent advancements related to generative AI are broadening the gap between the pros and cons of the technology even further, and underscore the need to act towards a common denominator.

Throughout this course, we will debate how balance between the pros and cons of the technology can be achieved. This will be done through unpacking the main pillars of AI, such as fairness, privacy, transparency and accountability, both from a CS perspective and from a legal/ policy perspective. The literature about each one of those components is on the rise both in the CS/ ML domains, and in the social sciences; and even binding laws are starting to enter into force. In each session, we will dive into one of the components, discuss from a theoretical perspective what are the considerations that each domain is adding to the equation, and demonstrate how practically, using real life case studies, we can bridge the gap and merge them into a solution that addresses both. A special focus will be given to the need to address contextual considerations and the implementation of AI in different domains such as healthcare, finance, and criminal justice.

Sessions overview

Session 1. Introduction- artificial intelligence, between technical definitions and policy definitions

The goal of this introductory session is to discuss the difference between the way the technical community and the legal/ policy community perceive terms such as AI and ML. we will discuss how the field of AI was developed, where it is standing now, and how it is expected to look like in the future, both from a technical and policy perspectives. In addition, we will introduce the latest work on AI governance and discuss what are the solutions that policy makers are considering in order to minimize the risk of AI systems, solutions ranging from soft law instruments such as guidelines, recommendations and codes of conducts, to legally binding laws like the EU AI Act, the Digital Marketing Act and the Digital Services Act.

Required readings:

1. Dennis Redeker, Lex Gill & Urs Gasser, Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights, 84 INT'L COMMUNICATION GAZETTE 302, (2018).
2. Peter Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo,, Defining AI in Policy versus Practice, AIES 20: Proceedings of the AAAI/ACM Conference on AI, Ethics and Society.
3. The full text of the UNESCO Recommendation on the Ethics of Artificial Intelligence

4. Jess Whittlestone, Rune Nyrup, Anna Alexandrova and Stephen Cave, The Role and Limits of Principles in AI Ethics: Towards A Focus on Tensions, AIES19 Proceedings of the AAAI/ACM Conference on AI, Ethics and Society,
5. Stanford HAI, Analyzing the European Union AI Act: What Works, What Needs Improvement, July 2023
6. International Association of Privacy Professionals, Contentious Areas in the EU AI Act Trilogues, August 2023
7. Alex Engler, The EU AI Act will have A Global Impact but Limited Brussels Effect, June 2023

Additional Readings:

1. Jonas Schuett , Defining the Scope of AI Regulation, (2019)
2. A general introduction to the OECD Principles on Artificial Intelligence,
3. Fjeld Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, Principled Artificial Intelligence: Mapping consensus in Ethical and Rights-Based Approaches to Principles for AI, Berkman Klein Center for Internet and Society (2020).
4. Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, Thomas Steinke, & Salil Vadhan, BRIDGING THE GAP BETWEEN COMPUTER SCIENCE AND LEGAL APPROACHES TO PRIVACY, 31:2 Harvard Journal of Law & Technology Volume 31:2 (Spring 2018).

Session 2. Fairness and nondiscrimination- technical perspectives

In this session we will unpack one of the main principles of AI, fairness and nondiscrimination. This principle is included in all documents, codes of conduct, and guidelines for governing AI. In addition, technical interest in this field expanded significantly, after stories about biased and discriminatory algorithms broke up in the media. In this class we will concentrate on the technical groups of fairness (individual fairness, group fairness and causal reasoning); and the sub definitions of fairness in each group. Among the specific notions that will be covered: fairness through awareness, statistical parity, equalized odds, and calibration. We will discuss the difference between the definitions, the benefits of choosing one over the other and their limitations.

Required Readings:

1. Moritz Hardt, Eric Price, Nathan Srebro (2016). Equality of Opportunity in Supervised Learning. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2. Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Dagstuhl, Germany.
3. A.K. Menon, RC.. Williamson (2018). The Cost of Fairness in Binary Classification. Conference on Fairness, Accountability, and Transparency (FAT 2018). Proceedings of Machine Learning Research 81:1–12.
4. Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application Vol.8:141–63.
5. Luca Oneto and Silvia Chiappa (2020). Fairness in Machine Learning. In: Oneto L., Navarin N., Sperduti A., Anguita D. (eds) Recent Trends in Learning From Data. Studies in Computational Intelligence, vol 896. Springer.
6. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi (2019). A Flexible Approach for Fair Classification. Journal of Machine Learning Research, 20(75):1−42.
7. Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data. 2017 Jun;5(2):153-163.

Additional readings:

1. Simon Caton and Christian Haas (2023). Fairness in Machine Learning: A Survey. ACM Comput. Surv. Just Accepted (August 2023).
2. Silvia, C., Ray, J., Tom, S., Aldo, P., Heinrich, J., & John, A. (2020). A General Approach to Fairness with Optimal Transport. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 3633-3640.

Session 3. Fairness and non-discrimination- legal perspectives

In this session we will focus on the legal and policy aspects of the technical definitions. In some domains there are specific legal mechanisms that bind the meaning of fairness and instruct what can or cannot be done. We will introduce some of those mechanisms, such as the disparate impact and disparate treatment theory in the US, we will discuss if those mechanisms are sufficient and whether they can be computed into a mathematical equation. In addition, we will return to the technical definitions discussed in the previous session, and assess to which policy domain they can be suitable.

Required readings:

1. Doaa Abu-Elyounes, "Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness," University of Illinois Journal of Law, Technology & Policy 2020, no. 1 1-54 (Spring 2020)

2. Thomas B. Nachbar, "Algorithmic Fairness, Algorithmic Discrimination," Florida State University Law Review 48, no. 2: 509-558 (Winter 2021)
3. Overview of Constitutional Requirements in Race-Concious Affirmative Action Policies in Education, The civil Rights Project, Harvard University
4. Daniel E. Ho, and Alice Xiang, Affirmative Algorithms: The Legal Ground for Fairness as Awareness, The University of Chicago Law Review Online, (2020).
5. Sandra Wachter, Brent Mittelstadt and Chris Russell, Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law (2021).

Additional readings:

1. The movie Codded Bias on Netflix
2. Reuben Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," Machine Learning Research 81:1-11 (2018).
3. Jon Elster, "Fairness and Norms," Social Research, Vol. 73, No. 2 (Summer 2006).
4. Margaret Kramer, A Timeline of Key Supreme Court Cases on Affirmative Action, The New York Times, 03/2019
5. Alice Xiang and Daniel E. Ho, From Affirmative Action to Affirmative Algorithms: The Legal Challenges Threatening Progress on Algorithmic Fairness, The Partnership on AI, 11/2020
6. Cass R. Sunstein, "Two Conceptions of Procedural Fairness," Social Research, Vol. 73, No. 2 (Summer 2006), pp. 619-646.
7. Solon Barocas, Andrew D. Selbst, Big Data's Disparate Impact, 104 California Law Review, 671 (2016).
8. Mark MacCarthy, "Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms, "48 CUMBER. L. REV. 1 (2017).

Session 4: Fairness and nondiscrimination- practical examples

After covering the core legal and technical theories, in this session we will discuss practical examples implementing the theory. We will learn and debate how fairness can be implemented in domains like criminal justice, healthcare, finance etc. The goal here is to dive deeper into the concepts we studied, and apply them on real life examples. Students who prepared critical reflection on the topic of fairness will be asked during the class to share from their work. In this class we will also discuss some technical and policy oriented tools for assessing fairness, and debate their utility.

Required reading:

- Select an article from the list of additional readings of previous classes that you have not read before and go through it.

Session 5. Privacy and data protection- legal perspectives

Building on the technical knowledge we acquired in the previous session, in this class we will examine the legal framework related to privacy and data protection. We will start with a general discussion about different notions of privacy in different places, mainly Europe and the US. Next, we will focus on the GDPR and discuss the changes that the regulation introduced and their impact on innovation and algorithmic developments, as well as the link between data protection and AI. Other issues that will be discussed depending on time constrains are data sovereignty and data sharing.

Required readings:

1. Julie E. Cohen, "What Privacy Is For," 126 HARV. L. REV. 1904 (2013)
2. Kobbi Nissim, et al. "Bridging the Gap between Computer Science and Legal Approaches to Privacy," 31 Harvard Journal of Law and Technology, 687 (2018),
3. Chris Jay Hoofnagle, Bart van der Sloot and Frederik Zuiderveen Borgesius, "The European Union general data protection regulation: what it is and what it means", 28 Information & Communications Technology Law, No. 1 65-98 (2019)
4. Whitman, James Q., "The Two Western Cultures of Privacy: Dignity versus Liberty", 113 The Yale Law Journal, no. 6 1151–1221 (2004).

Additional readings

1. Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, pp. 55-61 (PublicAffairs, 2019) (right to be forgotten)
2. Sandra Wachter," Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI"
3. Daniel J. Solove, "Conceptualizing privacy", 90 California Law Review, No. 4 1087-1055 (July 2002) (to read until page 1126)

Session 6. Privacy and data protection- technical perspectives

Harming the right to privacy and abuse of personal data is one of the main considerations in using AI algorithms and optimizing them. In this session we will explore different methods for preserving privacy. Those methods include differential privacy and cryptography.

Required Readings:

1. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference, pages 265-284. Springer, 2006.
2. Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 202-210, 2003.
3. Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), pages 94-103. IEEE, 2007.
4. Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In Proceedings of the 16th ACM conference on Computer and communications security, pages 534-544, 2009.
5. Rubinstein, Benjamin I. P., Peter L. Bartlett, Ling Huang, and Nina Taft. 2012. Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning. *Journal of Privacy and Confidentiality* 4 (1).
6. Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate (2011). Differentially Private Empirical Risk Minimization. Journal of Machine Learning Research, 12:1069-1109.

Session 7. Privacy and data protection- practical examples

In this session we will put together the accumulated knowledge we acquired in the last two sessions, and think how the legal and technical requirements play out in different policy domains where the privacy of individuals might be jeopardized. We will discuss some cases where advance privacy preserving techniques are applied in policy, like the use of differential privacy by the U.S. Census Bureau. We will also test practical tools for enhancing privacy, like privacy impact assessment tools.

Required readings

- Select an article from the list of additional readings on privacy and go through it.

Session 8. Transparency and explainability – legal perspective

In this session we will discuss different approaches to achieve explainability, from a legal perspective. We will learn about the difference between interpretable algorithms and non-interpretable algorithms, and the difference between transparency ex-ante and transparency ex-post including different auditing methods. We will also discuss the tradeoff between transparency and accuracy, and how balance between the two can be achieved. We will

match between different transparency methods and their relevant policy domains, and become familiar with legal concepts such as due process and fair trial; as well as the legal limitations that demand a certain form of transparency.

Required readings:

1. Doshi-Velez, Finale, and Mason Kortz, "Accountability of AI Under the Law: The Role of Explanation", Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper (2017)
2. European Commission Report on Safety and Liability Implications of AI, the Internet of Things and Robotics, (February 2020)
3. Ziosi, Marta and Mökander, Jakob and Novelli, Claudio and Casolari, Federico and Taddeo, Mariarosaria and Floridi, Luciano, The EU AI Liability Directive: Shifting the Burden From Proof to Evidence, available on SSRN (June 2023)
4. Seng, Daniel Kiat Boon, Artificial Intelligence and Information Intermediaries, NUS Centre for Technology, Robotics, Artificial Intelligence & the Law Working Paper 21/01, NUS Law Working Paper No. 2021/018, (July 2021)
5. Aina Turillazzi, Mariarosaria Taddeo, Luciano Floridi & Federico Casolari, "The digital services act: an analysis of its ethical, legal, and social implications", 15 Law, Innovation and Technology, 1, 83-106 (2023)

Additional readings

1. Aziz Z. Huq, Constitutional Rights in the Machine Learning State, 105 Cornell. L. Rev. (2020)
2. Zachary C. Lipton, "The Mythos of Model Interpretability," 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016, New York, NY) (Mar. 6, 2017)
3. Andrew D. Selbts and Solon Barocas, "The Intuitive Appeal of Explanable Machines," 87 Fordham L. Rev. 1085 (2018)

Session 9. Transparency and explainability – technical perspective

In this session we will discuss the technical aspects pertaining to explainability and transparency, and introduce different methods for ensuring explainability.

Required readings

1. Berk Ustun and Cynthia Rudin. Methods and Models for Interpretable Linear Classification. arXiv:1405.4047, (2014)
2. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144 (2016).
3. Vaishak Belle and Ioannis Papantonis. Principles and Practice of Explainable Machine Learning. Frontiers in Big Data, Sec. Data Mining and Management, Volume 4 – (2021)
4. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777 (2017).
5. Cynthia Rudin. Chaofan Chen. Zhi Chen. Haiyang Huang. Lesia Semenova. Chudi Zhong. "Interpretable machine learning: Fundamental principles and 10 grand challenges." Statist. Surv. 16 1 - 85 (2022).
6. Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, Ed Snelson. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. Journal of Machine Learning Research, vol. 14(101):3207−3260 (2013).
7. Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, Chirag Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. arXiv:2010.10596 (2020).

Additional readings

1. Senetaire, H.H.J., Garreau, D., Frellsen, J. & Mattei, P. Explainability as statistical inference. Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research 202:30584-30612 (2023)
2. Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, Sanparith Marukatat. Counterfactual Mean Embeddings. Journal of Machine Learning Research, vol. 22(162):1−71, 2021.

Session 10. Generative AI +the Impact of AI on the Environment

This session will be divided into two parts, in the first part we will discuss generative AI, and in the second part we will discuss the impact of AI on the environment (see description below for each one of the parts).

The release into the public domain and massive growth in the user base of artificial intelligence (AI) foundation models for text, images, and audio is fueling debate about the

risks they pose to work, education, scientific research, and democracy, as well as their potential negative impacts on cultural diversity and cross-cultural interactions, among other areas. In addition, such concerns are also reviving the debate about the need to govern the technology. Procedural framework to address and mitigate such concerns, including via effective governance models and tools such as ethical impact assessment and complementary approaches such as ethics by design or research ethics committees are needed. Unlike previous sessions where our focus was on a specific principle, now we will explore the significance of a technology itself, and assess how it would impact the principles.

With regards to the environment, on one hand, AI can strengthen climate predictions, enable smarter decision-making for decarbonising industries from building to transport, and work out how to allocate renewable energy. On the other hand, large machine learning systems consume a massive amount of resources, in fact, a Study projects that global carbon footprint from ICT will be equivalent to half of transportation's current level by 2040

In this part of the class, we will apply the principles we studied thus far on the domain of environmental protection. We will study the laws and policies that govern this domain. We will also examine few examples of technological solutions and assess their role in mitigating or exacerbating the problem.

Required readings:

1. Claudius Gros et al., Generative AI: A Concise Primer for Non-Experts, Institute for European Policy Making June 2023
2. UNESCO, Foundational Models such as ChatGPT through the Prism of the UNESCO Recommendation, June 2023
3. Luciano Floridi, AI as Agency Without Intelligence: on ChatGPT, Large Language Models and other Generative Models, Philosophy and Technology, 2023
4. Melissa Heikkila, AI Language Models are Rife with Political Biases, MIT Technology Review August 2023
5. Dylan Walsh, The Legal Issues Presented by Generative AI, MITSloan, August 2023.
6. Philipp Hacker, Regulating ChatGPT and other Large Generative AI Models, Proceedings of the 2023 ACM Conference on Fairness, Accountability and Transparency, June 2023

Evaluation

- Critical analysis of reading – 30%: the course is designed around three main blocks, fairness, privacy and transparency. You will be required to choose one of the three topics and within that block to choose one legal and one technical reading and to write a critical analysis about the chosen piece. You can choose whether to write the legal and technical analysis separately or combined, in any case, you should try to link between the two to the extent possible. Your analysis should include a brief summary of the paper, and a critical reflection. If you are writing a separate analysis for the legal and technical papers, each one should be about 800 words, if it is a combined analysis, it should be about 1500 words.
- 2 technical assignments 30%.
- Final project, mock trial 40%, in a group, you will be asked to simulate a court trial, you will pick a case study involving automation/ algorithm deployed in a certain domain. The group will be split into 2, where one side will present the plaintiffs (the party bringing forward the lawsuit), and the other side will present the defendant (the party that is being sued). You will be asked to apply the concept that we studied during the semester. The final projects will be presented during a special session that will be held on January 10.
- Active and meaningful participation would grant you a bonus to the grade