Indiana University Luddy School of Informatics, Computing, and Engineering

AI Topics Course: Artificial Intelligence Safety and Trust

Course listed as "Intro to AI and Infrastructure" ENGR E599, CSCI B659, INFO I590

Lecture: TTh 9:25-10:40 am
Location: Hodges Hall 2046 and/or via Zoom

Instructor: Beth Plale, McRobbie Professor of Computer Engineering, (812) 855-4373, plale@indiana.edu

I. Course Description

Artificial Intelligence (AI) today is enabling new advancements in AI and new applications and uses in science and society. Application of AI is raising concerns about fairness, autonomous decisions by machines, job loss through automation, accountability in safety-critical AI applications. This course raises student awareness of the concerns, and provides directions forward that originate from three sources: i) the individual developer or researcher and their organization, ii) the regulatory framework and professional societies, and iii) improvements to AI technology and its cyberinfrastructure.

The course is organized into four modules as follows:

- Module 1: AI Dilemmas and Scope: societal-facing dilemmas of Artificial Intelligence include issues of explainability, fragility, fakes, and bias.
- Module 2: AI Developer/researcher toolkit: techniques and conceptual frameworks for bringing about more responsible AI through their own efforts, including an ethical framework.
- Module 3: Regulatory and professional organizational influences on responsible AI.
- Module 4: AI Technology as solution: How technology itself can manifest more responsible AI.

II. Learning Objectives

After successfully completing the course, students will be able to (SWAT):

- Identify major society-facing dilemmas of AI including issues of explainability, fragility, fakes, and bias
- Bring an ethical framework to reasoning about ethical issues involving data about people
- Explain the importance of, and apply recommended improvements to, the reproducibility and replicability of their research or development work
- Apply Professional Society statements in algorithmic accountability and transparency in their work
- Attribute AI safety functions to various stakeholders: regulatory agencies, community organizations

- Explain AI safety in context of large scale, operationalized artificial general intelligence (AGI) company product
- Consciously apply ethical frameworks in reasoning about AI issues
- Apply concepts of fairness, transparency, and accountability
- Apply concepts of reproducibility and FAIR data
- Reason about role of national, international, and funder AI policies in future of AI
- Locate and apply techniques for how AI software and data can be improved to be more responsive to policies and inherent biases.
- Has demonstrated ability to tie the three perspectives (applications developer, public policy, AI software/data improvements) together in an assessment of progress on some dimension.

III. Reading List

Module 1: AI Risks

- *AI, Explain Yourself,* Communications of the ACM, Nov 2018, Vol. 61, No. 11
- *Dilemmas of Artificial Intelligence*, Peter J. Denning and Dorothy E. Denning, Communications of the ACM, March 2020, Vol 63, No. 3
- *Slaughterbots*, a documentary by Stuart Russell, CS professor of UC Berkeley, 2017 https://www.youtube.com/watch?v=9CO6M2HsoIA
- *Why we should ban lethal autonomous weapons*, Future of Life Institute documentary by top AI researchers. 2019. https://www.youtube.com/watch?v=LVwD-IZosJE

Module 2: Researcher/Practitioner Developer Toolbox

- Algorithmic Accountability: Designing for Safety. Ben Shneiderman, Invited talk at Radcliffe Institute for Advanced Study, Harvard University, May 2018. https://www.radcliffe.harvard.edu/video/algorithmic-accountability-designing-safety-ben-shneiderman
- Bit by Bit: Social Research in the Digital Age, Ch 2 and 6, Matthew J. Salganik, Princeton University Press. 2018.
- ACM Statement on Algorithmic Transparency and Accountability, https://www.acm.org/articles/bulletins/2017/january/usacm-statement-algorithmic-accountability
- Reproducibility and Replicability in Science, Chs. 1-5, National Academies of Science, Engineering, and Medicine, 2019
- Estimating the deep replicability of scientific findings using human and artificial intelligence, Yang Yang, Wu Youyou, and Brian Uzzi, PNAS, May 2020.

Module 3: Role of Regulator and Professional Society

- IEEE P7003TM Standard for Algorithmic Bias Considerations, Koene, Ansgar et al., ACM/IEEE International Workshop on Software Fairness, 2018.

- P7003 - Algorithmic Bias Considerations. Primary web site is IEEE P7003[1] but contains little information at present.
- Ethically Aligned Design, Second Ed., IEEE https://standards.ieee.org/news/2017/ead_v2.html
- Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight, Ben Shneiderman, PNAS Nov 29, 2016 113(48), doi.org/10.1073/pnas.1618211113

Module 4: Technological Manifestation

- *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, Miles Brundage and Shahar Avin et al., 2020, arXiv, arXiv:2004.07213v2
- XAI, Guide towards algorithmic explainability in machine learning, Alejandro Saucedo, PyData London 2019. https://www.youtube.com/watch?v=vq8mDiDODhc, https://github.com/EthicalML/xai
- Programming Your Way to Explainable AI @ O'Reilly AI NY 2017 Mark Hammond, Bonsai. https://www.youtube.com/watch?v=Um7grgYdBQQ
- How do we make AI Fair? Maya Gupta, Google AI, SysML 2019. https://www.youtube.com/watch?v=oCSmyzIBqZI
- Snapshot of the Frontiers of Fairness in Machine Learning. Alexandria Chouldechova and Aaron Roth, Communications of the ACM, Vol. 63, No 5, May 2020 pp. 82-89
- Industry-Scale Knowledge Graphs: Lessons and Challenges, Noy, Natasha et al., Communications of the ACM, Vol 62, No 8, Aug 2019
- Schema.org: Evolution of Structured Data on the Web, R. V. Guha, Dan Brickley, Steve Macbeth, Communications of the ACM, Feb 2016, Vol 59, No. 2

Tying it together

Tying it together. In the final 2 weeks of the semester, students will each have 20 or so minutes to present and discuss their report which has been written on a topic that is comprehensive over the content of the semester

[1] https://standards.ieee.org/project/7003.html