TEC Technology Ethics Center

UNIVERSITY OF NOTRE DAME

TEC 30114 Auditing AI: An Introduction
Fall 2023

Teachers: Cameron Kormylo

Class Meets: 2X per week (75-minute class periods)

Office Hours: _____

Course Overview and Objectives

Course Overview

As artificial intelligence (AI) grows increasingly pervasive in society, it is essential that we develop an understanding of how AI systems work. A vital part of this understanding is a careful consideration of various risks (e.g., the presence of bias, a lack of transparency, regulatory compliance) when AI systems are designed and deployed in real-world settings.

To understand and address these concerns, this course introduces students to the fundamentals of AI auditing — the practice of evaluating and improving the ethics of AI systems. Through a combination of interactive discussions and technical lab sessions, students will develop an auditing “toolkit”. This toolkit includes both theoretical and technical concepts, especially relevant for the increasingly interdisciplinary teams of the modern workforce. Students will work on group case assignments as “audit committees” that reflect the needs of a variety of stakeholders (e.g., developers, managers, investors, users). Groups will identify and discuss potential concerns or risks associated with AI systems as well as develop recommendations to address them.

Overall, the course aims to provide an interdisciplinary and hands-on introduction to AI auditing, allowing students to gain insights into the opportunities and challenges associated with the design and deployment of AI systems that minimize societal risk and increase their effectiveness.

Course Objectives

In this course, you will:

- Develop an understanding of the concerns associated with the ethical design and deployment of AI systems in the real world.
- Explore key theoretical and technical concepts in the emerging research on AI audits and examine their underlying assumptions.
- Critically engage with these concepts to assess a range of AI systems developed from real-world datasets and examples.
- Identify potential risks associated with AI systems and possible recommendations to mitigate these risks.
- Work in interdisciplinary teams that represent different stakeholders to explore a holistic and sustainable approach to auditing AI systems.
- Develop the capacity to consider different risks and mitigation steps when designing, deploying, or working with AI systems in your future research and practice.

Required Materials

Textbook: None.

Articles: All readings are provided on Canvas.

Video Series: A video series developed specifically for this course will be made available for students. The entirety of the series will be assigned throughout the course. Video topics are listed below as well as in the course schedule on their assigned dates.

1. A Technical Introduction to Auditing AI
2. Traditional Measures of Performance: Accuracy, Precision, Recall, & Beyond
3. Auditing the Entirety of the AI Lifecycle
4. Synthetic Data & its Role in AI Audits
5. Data Collection & Selection: Identifying Bias At the Outset
6. Removing the Black Box: Auditing for Interpretation
7. Testing for Fairness: A/B Testing & Modern Metrics
8. Privacy After Deployment: Using Adversarial Testing to De-Anonymize AI Outputs

Software: For the technical portion of the course, we will utilize various toolkits and packages in Python. Prior experience with Python is not required for this course. For students who are new to Python, I recommend you install Anaconda as we will be utilizing Jupyter for in class demonstrations. The following is an easy to follow guide on installing and using Jupyter through Anaconda.

https://www.dataquest.io/blog/jupyter-notebook-tutorial/

If you already have a preferred IDE (Atom, VSCode, etc.), feel free to use that instead.

Course Format, Assignments, Evaluation and Grading

Format

For the first half of the course (Wk1 - Wk7), we will combine (i) a discussion session to explore key theoretical concepts in the assigned readings and (ii) a lab session where students will learn various technical auditing methods and will be given the opportunity to practice these methods on a variety of datasets. The technical concepts have been selected to go together with the theoretical concepts covered in that day's discussion session. For the second half of the course (Wk8 - Wk14), students will mainly work in groups to analyze cases involving real-world and/or hypothetical AI systems. Each group will consist of interdisciplinary students acting as a mini audit committee, reflecting the multiple stakeholders who participate in auditing real-world AI systems. For each week's case, groups will identify a key concern or risk (be it social, organizational, or technical) and recommend steps to address it. All groups will briefly present their concerns and proposed solutions, based on which we will have a discussion exchanging multiple points of view. This course seeks to open up a collaborative learning space in which students can creatively explore different ways of analyzing, assessing, and improving AI systems in a variety of real-world contexts.

Performance Measurements

- 20% Reflection notes (10% x 2) [individual]
- 15% Midterm exam [individual]
- 30% Case presentation + report (5% x 6 cases) [group]
- 20% Final case presentation + report [group]
- 15% Participation [individual]

Reflection notes [individual] (20% = 10% x 2)

Each student is expected to submit two written reflection notes (1-2 pages), one before and one after the midpoint of the semester. For these notes, students should find a relevant real-world example of a deployed or proposed AI system. They should then analyze and reflect on the example by drawing on the concept(s) and/or case(s) discussed during the course. The notes will be graded by the extent to which you engage with concepts and cases from the course in your analysis, as well as the originality of your reflections.

Reflection Note Rubric	No credit	Partial credit	Full credit
Critically engaging with	No course	Mentioned course	Analysis based on

course material 50%	materials mentioned	materials, but with no/little analysis or not connected to the theme	course materials and connected to the theme
Offering original reflections 50%	No reflections offered	Reflections offered but not connected to course materials, or not original	Original reflections based on analysis of course materials

<u>In-class midterm exam [individual] (15%)</u>

Before we begin analyzing cases involving real-world and/or hypothetical AI systems, we will have an in-class midterm exam covering the key concepts discussed in the initial part of the course. While students will not be expected to produce original code or work with data for the exam, questions related to the application of technical methods will be included.

<u>Case presentation + report [group] (30% = 5% X 6)</u>

Students will work in groups to write and present short reports (1-2 pages) on six cases involving real-world or hypothetical AI systems (one case per week). For each case, groups will represent a different stakeholder and will be provided with a case synopsis, relevant articles, and a corresponding dataset and/or AI model. In the first half of each week's class session, groups will work to identify one concern or risk about the AI system that is relevant to the stakeholder group they represent. Additionally, groups will recommend potential remediation steps to address their concerns. In the second half of class, groups will briefly present their concerns and recommendations to the class (< 5 minutes). Students will then together exchange points of view from the various stakeholder perspectives in an interactive discussion. Groups are expected to submit a summary report (1-2 pages) of their presentation.

Case Rubric	No credit	Partial credit	Full credit
Presentation 50%	No presentation	Presented but does not identify a clear or valid concern or make clear or valid recommendations	Presented both the concern and recommendations clearly
Written report 50%	No written report	*Summarized only the concern or only the recommendations *Lacks readability	*Summarized both the concern and recommendations *Clearly written

<u>Final case presentation + report [group] (20%)</u>

Towards the end of the semester, you will be provided with a final audit case package (case synopsis + relevant articles + dataset/AI model). For this final case, you will be expected to bring in and synthesize perspectives of multiple stakeholders to produce a more comprehensive audit report. At the end of the course, your group will present your audit report (< 10 minutes) and submit a written report (6-8 pages).

Final case presentation + report rubric

Presentation 50%	Content / presentation skills / equal contribution
Written report 50%	Content / structure

Participation [individual] (15%)

Participation is a significant element of your grade for this course. Each student is expected to actively and respectfully participate in class discussions, both as an individual and as a group member. Evaluation will be based on in-class participation as well as a peer review.

Grading

Grading criteria for assignment of final course grade is based on:

Percentage of Course Points Earned	Grade Earned
93.0 % and above	A
from 90.0 % but less than 93.0 %	A-
from 87.0 % but less than 90.0 %	B+
from 83.0 % but less than 87.0%	B
from 80.0 % but less than 83.0%	B-
from 77.0 % but less than 80.0 %	C+
from 73.0 % but less than 77.0 %	C
from 70.0 % but less than 73.0 %	C-
from 60.0 % but less than 70.0 %	D
less than 60.0%	F

The students are expected to adhere to the University Honor Code, *Student Guide to Academic Code of Honor* (www.nd.edu/~hnrcode). Any violations of the Honor Code will be referred to the appropriate committee.

Course Schedule

	Topic	Reading(s) / Video(s)
Wk1	**AI Audits** **Discussion 1:** A Brief History	

	of AI Audits & Brainstorming an Audit of ChatGPT **Lab Session 1:** Methods Overview and Traditional Performance Measures	● S. Michael Gaddis. 2018. An introduction to audit studies in the social sciences. In S.M. Gaddis (ed.), *Audit studies: behind the scenes with theory, method, and nuance*, Springer International, Cham, 3-44. ● Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. *Auditing algorithms: understanding algorithmic systems from the outside in*, Now Publishers, Boston and Delft, 3-23. ● https://hbr.org/2022/12/chatgpt-and-how-ai-disrupts-industries ● https://openai.com/blog/chatgpt/ ● https://www.dataquest.io/blog/jupyter-notebook-tutorial/ ● Video 1 - A Technical Introduction to Auditing AI ● Video 2 - Traditional Measures of Performance: Accuracy, Precision, Recall, & Beyond
Wk2	**Audit Frameworks** **Discussion 2:** Auditing Frameworks **Lab 2:** The AI Lifecycle & Generating Synthetic Data	● Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 33-44. ● Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing algorithms: understanding algorithmic systems from the outside in, Now Publishers, Boston and Delft, 24-50. ● Kirsten Martin. 2019. Designing ethical algorithms. *MIS Quarterly Executive* 18, 2, 129-142. ● Video 3 - Auditing the Entirety of the AI Lifecycle ● Video 4 - Synthetic Data & its Role in AI Audits
Wk3	**Data** **Discussion 3**: The Challenges of Real-World Data	● Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. "Bringing the people back in: contesting benchmark machine learning datasets." *arXiv preprint arXiv:2007.07399*. ● Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. Communications of the ACM 64, 12, 86-92. ● Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 560-575.

	Lab 3: Identifying Patterns and Bias in Data	● Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 1, 665-673. ● Video 5 - Data Collection & Selection: Identifying Bias At the Outset
Wk4	**Explainability** **Discussion 4:** Unblackboxing Algorithms / Explainability for What & Whom? **Lab 4:** Model Interpretation	● Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 220-229. ● Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: an overview of interpretability of machine learning. In IEEE Conference on Data Science and Advanced Analytics (DSAA), 80-89. ● Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion 58, 82-115. ● Susan Athey. 2017. Beyond prediction: using big data for policy problems. *Science* 355, 6324, 483-485. ● Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence 1, 5, 206-215. ● Video 6 - Removing the Black Box: Auditing for Interpretation
Wk5	**Fairness** **Discussion 5:** Defining Fairness **Lab 5:** Fairness Metrics and A/B Testing	● Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys 54, 6, 1-35. ● Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 39-48. ● Video 7 - Testing for Fairness: A/B Testing & Modern Metrics
Wk6	**Privacy** **Discussion 6:** Differential Privacy and AI	● Tianqing Zhu, Dayong Ye, Wei Wang, Wanlei Zhou, and S. Yu Phillip. 2020. More than privacy: Applying differential privacy in key areas of artificial intelligence. In IEEE Transactions on Knowledge and Data Engineering, 34(6), 2824-2843. ● Huaxin Li, Haokin Zhu, Suguo Du, Xiaohui Liang, and Xuemin Shen. 2018. Privacy Leakage of Location Sharing in Mobile

	Lab 6: De-Anonymization & Adversarial Testing	Social Networks: Attacks and Defense. In IEEE Transactions on Dependable and Secure Computing, 15(4), 646-660. ● Video 8 - Privacy After Deployment: Using Adversarial Testing to De-Anonymize AI Outputs
Wk7	**In-class exam**	
Wk8	**Case 1: Automated Recruiting** Overview of the case and dataset/model. Group presentation + discussion.	● Case synopsis (reference articles) ● Mona Sloane, Emanuel Moss, and Rumman Chowdhury. A Silicon Valley love triangle: hiring algorithms, pseudo-science, and the quest for auditability. *Patterns* 3, 2, 100425. ● Data/Model ● https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset *[*First reflection note due]*
Wk9	**Case 2: Healthcare** Overview of the case and dataset/model. Group presentation + discussion.	● Case synopsis (reference articles) ● https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html ● Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11, 745-750. ● Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1-14. ● Data/Model ● https://www.kaggle.com/datasets/fedesoriano/covid19-effect-on-liver-cancer-prediction-dataset
Wk10	**Case 3: Predicting the Property Market** Overview of the case and dataset/model. Group presentation + discussion.	● Case synopsis (reference articles) ● https://www.wired.com/story/zillow-taps-ai-improve-home-value-estimates/ ● https://edition.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html ● Data/Model ● https://www.kaggle.com/code/erick5/predicting-house-prices-with-machine-learning

Wk11	**Case 4: Psychological Targeting** Overview of the case and dataset/model. Group presentation + discussion.	● Case synopsis (reference articles) ● Erik Hermann. 2022. Psychological targeting: nudge or boost to foster mindful and sustainable consumption? *AI & Society,* 1-2. ● https://aeon.co/essays/dreams-are-a-precious-resource-dont-let-advertisers-hack-them ● Data/Model ● https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets
Wk12	**Case 5: Saving Wildlife** Overview of the case and dataset/model. Group presentation + discussion.	●Case synopsis (reference articles) ● Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1-12. ● HBR - SMART: AI and ML for wildlife conservation ● Data/Model ● https://projects.iq.harvard.edu/files/teamcore/files/2017_7_teamcore_ibmjournal_accepted.pdf
Wk13	**Case 6: AI Character-Driven News** Overview of the case and dataset/model. Group presentation + discussion.	●Case synopsis (reference articles) ● https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html ● Data/Model ● https://www.kaggle.com/competitions/deepfake-detection-challenge/overview *[*Second reflection note due]*
Wk14	**Final Presentations + Discussion**	*[*Participation peer review]* *[*Final audit report due]*