Carnegie Mellon University

94885 Responsible AI - Principles, Policies and Practices

Meeting Days, Times, Location: TBD

Semester/Year: Spring 2024

Units: 6**, Section(s):** A

Instructor information

Name	Prof Anand Rao
Contact Info	anandr2@andrew.cmu.edu
Office location	TBD
Office hours	TBD

Course Description

As the world rapidly embraces Artificial Intelligence, the potential for both benefit and harm escalates. This course, "Responsible AI: Principles, Policies, and Practices," navigates the complexities of responsible AI use. Our focus is on providing a detailed and practical understanding of the key risks and harms traditional and generative AI can pose, the principles guiding ethical use of AI, and the intricacies of how these harms manifest themselves in the AI lifecycle. This course places a strong emphasis on bias, fairness, transparency, explainability, safety, security, privacy, and accountability, demystifying these foundational concepts and highlighting their relevance in the end-to-end AI life cycle.

Delve into the regulatory landscape of AI as we dissect policymaking worldwide and scrutinize responsible AI frameworks adopted by leading organizations. You'll gain valuable insight into the emerging standards, certifications, and accreditation programs that are guiding the responsible use of AI, Generative AI, and Large Language Models. Building on this knowledge, the course will help you understand the integral role of governance in AI and the pivotal role that various stakeholders play in this landscape.

Our unique approach combines theory with practical strategy, enabling you to develop a comprehensive operational plan for implementing responsible AI within an organization. The course culminates with the creation of a strategy and handbook tailored to the needs of an organization. Furthermore, we will equip you with the skills to communicate effectively, making a compelling case for implementing a responsible AI program. Several guest lectures from practitioners and policy makers, coupled with synthetic case scenarios will give you a window into how organizations and policy making bodies are advancing the responsible use of AI.

Whether you're a technology enthusiast or policy student, if you possess a basic understanding of data science and artificial intelligence, this course is a golden opportunity to immerse yourself in the riveting world of responsible AI. Join us as we explore, analyze, and operationalize Responsible AI from a vantage point that fuses ethical considerations with technical prowess.

Learning Objectives

1. Evaluate and categorize the key risks and harms associated wtih traditional and generative AI.
2. Critically assess and apply ethical principles and trade-offs in the use of AI technologies.
3. Identify and map the stages of the AI system lifecycle, pinpointing where risks and harms are most likely to manifest.
4. Develop and implement strategies to manage and mitigate issues of bias, fairness, transparency, explainability, safety, security, privacy, and accountability in AI systems.
5. Evaluate and compare global regulatory and policy frameworks related to AI, Generative AI, and Large Language Models.
6. Analyze and critique how various companies adopt and operationalizing Responsible AI frameworks.
7. Evaluate and differentiate between emerging standards, certifications, and accreditation programs in the field of responsible AI.
8. Create end-to-end and top-down AI governance models for AI, identifying the roles and responsibilities of different stakeholders.
9. Formulate and document a comprehensive strategy for operationalizing responsible AI within an organization.
10. Effectively articulate and present the rationale and mechanics of a responsible AI program to both technical and non-technical stakeholders within an organization.

Learning Resources

The following textbooks will be used as reference for the topics covered. Each topic will also have selected reading materials.

1. Responsible AI in the Enterprise by Adnan Masood, Heather Dawe, and Ehsan Adeli, Packt Publishing, July 2023.
2. Machine Learning for High-Risk Applications by Patrick Hall, James Curtis, and Parul Pandey, O'Reilly Media, Inc., April 2023.
3. Trustworthy AI by Beena Ammanath, Wiley, March 2022.

Assessments

The final course grade will be calculated using the following categories:

Assessment	Percentage of Final Grade
Class Participation	10%
Individual Assignment 1	20%
Individual Assignment 2	20%
Best 2 of 3 Quizzes	20%
Team Project Presentation	30%
Total	100%

- **Class Participation:** Class participation would be based on (a) Coming prepared to the class having read the pre-reads; (b) Meaningful contributions to the case discussions and insightful questions during the lectures.
- **Individual Assignments 1 and 2:** Individual assignments 1 and 2 will be based on answering the discussion questions in the industry-based synthetic scenarios that will be distributed at least one week before the assignments are due. See note on the use of generative AI tools in the Generative AI Guidance section.
- **Quizzes**: Three classroom/online quizzes will be administered during Weeks 2 to Week 6 of the course. Each quiz will be 10% of the total score and the best two quizzes will be taken for the final quiz assessment of 20%. Students are NOT allowed to use any AI tools or textbooks for the quizzes.
- **Team Project Presentation:** The final project will be a team presentation based on an industry-based synthetic scenario that will be distributed at the start of the course. The students will work as a team during the course and will make the final presentation to a panel of judges. There will be no final exam and the presentation will be conducted during the week of the exams.

Students will be assigned the following final letter grades, based on calculations coming from the course assessment section.

Grade	Percentage Interval
A+	98.0-100%
A	92.0-97.9%
A-	90.0-91.9%
B+	88.0-89.9%
B	82.0-87.9%
B-	80.0-81.9%
C+	78.0-79.9%
C	72.0-77.9%
C-	70.0-71.9%
D	50.0-69.9%
F	00.0-49.9%

Grading Policies

- **Late-work policy**: To encourage timely submissions and ensure fair and prompt grading for all students, assignments should be submitted by 11:59 PM on the due date. For those facing unforeseen circumstances, assignments may be submitted up to 24 hours late for up to 90% of the original grade, with incremental reductions thereafter. No assignments will be marked after 10 days.
- **Make-up work policy**: To maintain the integrity of the grading process while offering flexibility, there will be no make-up assignments or quizzes. However, students can miss one out of the three quizzes without

affecting the full score of 20% allocated for quizzes.

- **Re-grade policy**: To uphold the integrity of the assessment process, regrading will not be available. However, students are welcome to discuss the rationale for their grades during office hours to gain a better understanding of the assessment.
- **Attendance and/or participation policy**: To emphasize the value of class participation and active engagement in the learning process, attendance is mandatory and will be tracked via a sign-in sheet. Students have the flexibility to miss one class without affecting their class participation grade, as outlined in the Class Participation guidelines.

Course Policies

- **Academic Integrity & Collaboration**: Students are expected to strictly follow Carnegie Mellon University rules of academic integrity in this course. This means that unless otherwise specified, Individual assignments are to be the work of the individual student using only permitted material and without any cooperation of other students or third parties. It also means that usage of work by others is only permitted in the form of quotations and any such quotation must be distinctively marked to enable identification of the student's own work and own ideas. All external sources used must be properly cited, including author name(s), publication title, year of publication, and a complete reference needed for retrieval. The same work may not be submitted for credit in multiple courses. Violations will be penalized to the full extent mandated by the CMU policies. There will be no exceptions.
- **Use of Generative AI Tools**: We encourage students to explore the use of generative artificial intelligence (AI) tools, such as ChatGPT, for all individual assignments. Any such use must be appropriately acknowledged and cited, following the guidelines established by the APA Style Guide, including the specific version of the tool used. Submitted work should include the exact prompt used to generate the content as well as the AI's full response in an Appendix. Because AI generated content is *not* necessarily accurate or appropriate, it is each student's responsibility to assess the validity and applicability of any generative AI output that is submitted. You may not earn full credit if inaccurate, invalid, or inappropriate information is found in your work. Deviations from these guidelines will be considered violations of CMU's academic integrity policy.
- **Disabilities**: If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.
- **Student wellness**: As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at http://www.cmu.edu/counseling . Support is always available (24/7) from Counseling and Psychological Services: 412-268-2922.
- **Diversity:** It is my intent that students from all diverse backgrounds and perspectives be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength, and benefit. It is my intent to present materials and activities that are respectful of diversity: gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally or for other students or student groups.

Course Outline

This mini-course is planned around thirteen sessions of 1 hour 20 minutes apiece and will be taught in Hamburg Hall 2008.

- **Lecture 1: Introduction & Overview**
  - **Topics**
    - Introductions
    - Structure of the class
    - Expectations
    - Why study Responsible AI?
    - What will you get out of it?
  - **Readings**:
    - (40 minutes) [Masood] Chapter 1 on “Explainable and Ethical AI Primer” from Responsible AI in the Enterprise
    - (15 minutes) [Ammanath] Chapter 1 on “A Primer on Modern AI” from Trustworthy AI
- **Lecture 2: Murphy’s Law in AI: Case Study**
  - **Topics**
    - Case study discussion
    - AI harms
    - AI Risks
  - **Readings**:
    - (45 minutes) [Masood] Chapter 2 on “Algorithms Gone Wild” from Responsible AI in the Enterprise
  - **Advanced Readings:**
    - Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy by Cathy O’Neil
- **Lecture 3: AI Risk Management**
  - **Topics**
    - Five Views of AI Risk
    - Risk taxonomy
    - NIST AI Risk Management Framework
  - **Readings**:
    - (5 minutes) Five views of AI Risk: Understanding the darker side of AI by Anand Rao
    - (90 minutes) Artificial Intelligence Risk Management Framework, NIST
  - **Advanced Readings:**
    - (25 minutes) TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI by Andrew Critch and Stuart Russell.
- **Lectures 4&5: Bias & Fairness Case Study**
  - **Topics**
    - Bias and Fairness Case Study
    - Human Bias
    - Nature of Bias in AI
    - Bias vs Fairness
- **Lectures 4&5: Fairness Metrics and Tools**
  - **Topics**
    -

        - Fairness metrics
        - Tradeoffs in Fairness
        - Demonstration of Aequitas
        - Leading practices in promoting Fairness
    - **Readings**:
        - (20 minutes) [Ammanath] Chapter 2 on "Fair and Impartial" from Trustworthy AI
        - (120 minutes) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST Special Publication 1270
    - **Advanced Readings:**
        - [Masood] Chapter 8 on "Fairness in AI Systems with Microsoft Fairlearn" from Responsible AI in the Enterprise
        - [Masood] Chapter 9 on "Fairness Assessment and Bias Mitigation with Fairlearn and the Responsible AI toolkit" from Responsible AI in the Enterprise
        - [Hall] Chapter 4 on "Managing Bias in Machine Learning" from Machine Learning for High-Risk Applications
        - (25 minutes) Aequitas: A Bias and Fairness Audit Toolkit by Saleiro et. Al.,
        - What is fair when it comes to AI bias? by Anand Rao & Ilana Golbin
- **Lectures 6: Explainability & Interpretability Case Study**
    - **Topics**
        - Explainable and Interpretable Case Study
        - Transparency, Explainability, and Interpretability
        - Four Principles of Explainable AI
        - Five critical questions for Explainable AI
- **Lecture 7: Explainability & Interpretability Details**
    - **Topics**
        - Self-interpretable and post-hoc explanations
        - Model-Agnostic Explanations
        - Local Explanation – SHAP
        - Counterfactual Explanation
        - Explanation Evaluation
    - **Readings**:
        - [Masood] Chapter 3 on "Opening the Algorithmic Black Box" from Responsible AI in the Enterprise
        - [Ammanath] Chapter 5 on "Explainable AI" from Trustworthy AI
        - [Ammanath] Chapter 4 on "Transparent" from Trustworthy AI
    - **Advanced Readings:**
        - [Masood] Chapter 7 on "Interpretability toolkit and fairness measures" (Pages 175-185) from Responsible AI in the Enterprise
        - [Hall] Chapter 2 on "Interpretable and Explainable Machine Learning" from Machine Learning for High-Risk Applications
- **Lecture 8: Guest Lecture on Operationalizing Responsible AI**
- **Lecture 9: Safety, Security, and Privacy**
    - **Topics**
        - Overview of safety, security, privacy, resilience
        - Taxonomy of AI safety and Security
        - Adversarial attacks and mitigation
        - Model and data security
        - Monitoring and mitigating model drift

        - Privacy and privacy-enhancing technologies
    - **Readings**:
        - Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations by Alina Oprea and Apostol Vassilev, NIST AI 100-2e2023, March 2023.
        - [Masood] Chapter 4 on "Robust ML – Monitoring and Management" from Responsible AI in the Enterprise
        - [Ammanath] Chapter 3 on "Robust and Reliable" from Trustworthy AI
        - [Ammanath] Chapter 6 on "Secure" from Trustworthy AI
        - [Ammanath] Chapter 7 on "Safe" from Trustworthy AI
        - [Ammanath] Chapter 8 on "Privacy" from Trustworthy AI
    - **Advanced Readings:**
        - [Hall] Chapter 3 on "Debugging Machine Learning Systems for Safety and Performance" from Machine Learning for High-Risk Applications
        - [Hall] Chapter 5 on "Security for Machine Learning" from Machine Learning for High-Risk Applications
- **Lecture 10: Regulations, Standards, Certification**
    - **Topics**
        - Policies and regulations
        - Industry standards and professional bodies
        - Certifications
        - AI Audit
        - Responsible AI toolkits
    - **Readings**:
        - [Masood] Chapter 5 on "Governance, Audit, Compliance" from Responsible AI in the Enterprise
        - [Hall] Chapter 1 on "Contemporary Machine Learning Risk Management" from Machine Learning for High-Risk Applications
- **Lecture 11: Governance, Risk, Compliance in Enterprises**
    - **Topics**
        - Enterprise AI Governance
        - Role of Board and C-suite
        - Risk, Compliance, and Operational roles in Responsible AI
        - NIST AI RMF Playbook
    - **Readings**:
        - [Masood] Chapter 6 on "Enterprise AI starter kit for Fairness, Accountability, and Transparency" from Responsible AI in the Enterprise
        - [Ammanath] Chapter 9 on "Accountable" from Trustworthy AI
        - [Ammanath] Chapter 10 on "Responsible" from Trustworthy AI
        - [Ammanath] Chapter 11 on "Trustworthy AI in Practice" from Trustworthy AI
        - [Ammanath] Chapter 12 on "Looking Forward" from Trustworthy AI
- **Lecture 12: Guest Lecture on Standards and Policy Formulation**
    - **Topics**
        - Standards and Policy Formulation at NIST by Reva Schwartz, Research Scientist/Principal Investigator for AI Bias at National Institute of Standards and Technology (NIST)
- **Lecture 13: Summary & Reflections**
    - **Topics**
        - Summary of course

- Team project and final presentation logistics