Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Carnegie Mellon University

Responsible AI - Principles, Policies and Practices

Course Developer: Professor Anand Rao

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 Resources

The following textbook will serve as a primary reference for the topics discussed.

1. Trustworthy AI by Beena Ammanath, Wiley, March 2022.

Additionally, selected reading materials will be provided for each topic. Two supplementary books that delve into the course material in greater detail are:

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

Learning Objectives

Students should be able to:

- Understand and Apply Foundational Principles of Responsible AI: Identify and evaluate key ethical concepts such as bias, fairness, and transparency, and apply global AI regulatory frameworks to industry use cases.
- Assess Short- and Long-Term AI Risks to Stakeholders: Critically assess the risks and harms AI poses to different stakeholders (individuals, corporations, society), especially in high-impact sectors like healthcare and finance.
- Apply the NIST AI Risk Management Framework (RMF): Learn to apply the NIST AI RMF to assess risks in AI systems, focusing on security, fairness, transparency, and governance, and develop practical strategies for managing these risks across the AI lifecycle.
- Measure and Mitigate Bias and Fairness in AI Systems: Apply fairness metrics (e.g., demographic parity, predictive parity) and evaluate strategies for mitigating bias in AI systems, focusing on real-world applications in healthcare, finance, and criminal justice.
- Enhance AI Explainability and Transparency: Apply techniques such as LIME and SHAP to improve the explainability of AI models, balancing transparency with accuracy for diverse stakeholders in technical and non-technical roles.
- Evaluate and Address AI Privacy Risks: Analyze privacy risks in AI systems (e.g., data reconstruction, membership inference), and apply privacy-enhancing technologies (e.g., differential privacy, homomorphic encryption) to mitigate these risks.
- Ensure AI Safety, Robustness, and Reliability: Measure and manage AI safety risks, including adversarial attacks and model drift, by implementing adversarial training and stress testing, focusing on robustness and resilience across deployment environments.
- Develop and Apply AI Governance Tools: Apply governance tools, such as AI impact assessments and algorithmic audits, to ensure the ethical deployment of AI, and compare global approaches to AI governance (e.g., U.S., EU, Singapore).
- Implement Responsible AI Systems in Real-World Contexts: Develop governance frameworks using real-world case studies, integrating fairness, transparency, safety, and privacy practices to responsibly manage AI systems.

Assessments

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

Assessment	Percentage of Final Grade
Class Participation	10%
Six Quizzes (5% each)	30%
Individual Assignment 1	20%
Individual Assignment 2	20%
Team Assignment	20%
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.
- **Quizzes:** Six quizzes will be administered during the course – one each week. Each quiz will be 5% of the total score. Students are NOT allowed to use any AI tools or textbooks for the quizzes.
- **Team Projects:** There will be two individual projects and one team project. There will be no final exam and the presentation will be conducted during the week of the exams or the final lecture.

Course Schedule

Week	Theme	Learning Outcomes Addressed	Assignments Due
1	L1: Introduction & Overview	**Understand Responsible AI principles:** Introduce bias, fairness, transparency, and their relevance in AI governance. **Evaluate AI harms:** Assess the impact of AI harms on stakeholders in sectors like healthcare and finance. **Analyze real-world Cases:** Review real-world cases to explore risks associated with premature AI deployment **Case Study:** Responsible AI at tech firms	
	L2: AI Risk Management	**Mitigate AI risks:** Develop strategies to categorize and address risks in the AI arms race, using practical examples. **Apply risk management frameworks:** Use the NIST framework to assess risks and implement governance strategies across the AI lifecycle **Case Study:** Responsible AI at tech firms	(Quiz-1)
2	L3: Mapping and Measuring Bias and Fairness	**Understand bias and fairness in AI systems:** Compare statistical and social bias definitions using examples from healthcare and criminal justice to explore their impact on AI systems. **Evaluate types of bias in the AI lifecycle:** Identify and assess different types of bias (statistical, systemic, cognitive) in AI development, and analyze their effects on decision-making. **Apply fairness metrics to assess AI bias:** Use fairness metrics like demographic parity and equal opportunity with the Google PAIR tool to explore trade-offs in AI fairness. **Class Activity:** Google PAIR tool	

	L4: Managing Bias and Fairness	**Evaluate and apply bias mitigation strategies in AI systems:** Analyze and apply pre-processing, in-processing, and post-processing techniques to mitigate bias, assessing their effectiveness in different AI contexts. **Analyze legal and socio-technical challenges in fairness:** Apply fairness principles to evaluate legal issues like disparate impact and explore socio-technical challenges in sectors like criminal justice. **Assess fairness trade-offs in AI-driven decision-making:** Use the Courtroom Algorithm Game to critically evaluate fairness trade-offs, such as false positives vs. false negatives, in AI systems used for legal decisions. **Class Activity:** Courtroom Algorithm Game	(Quiz-2)
3	L5: Mapping and Measuring Explainability and Interpretabili ty	**Map explainability risks and suggest improvements:** Use the NIST AI Risk Management Framework to identify and evaluate explainability risks, proposing ways to address transparency and interpretability issues in AI systems. **Evaluate and compare explainability techniques:** Apply LIME and SHAP to assess and compare their effectiveness in improving explainability, focusing on their clarity for both technical and non-technical audiences	
	L6: Managing Explainability and Interpretabili ty	**Create strategies for managing explainability:** Develop stakeholder engagement and compliance strategies using the NIST AI Risk Management Framework, ensuring transparency while balancing accuracy and fairness in AI systems. **Propose solutions to balance accuracy and explainability trade-offs:** Analyze case studies like OptiClaim to evaluate trade-offs, and formulate recommendations that address the needs of both technical and non-technical stakeholders. **Case Study:** OptiClaim Solutions	(Quiz-3) (Individual Assignment 1)
4	L7: Privacy in AI: Mapping, measuring, and managing	**Map and assess privacy risks in AI systems:** Identify key privacy risks, such as data reconstruction and membership inference, by mapping them across the AI lifecycle using the NIST AI RMF. Apply this mapping to real-world AI systems to pinpoint vulnerabilities.	

		Evaluate and apply privacy-enhancing technologies: Analyze privacy-enhancing techniques like differential privacy and homomorphic encryption to assess their effectiveness in mitigating risks. Apply these methods to balance privacy, accuracy, and compliance in AI systems. **Develop strategies to manage privacy risks and failures:** Formulate strategies for managing privacy risks using the NIST AI RMF. Use the Cambridge Analytica case to evaluate failures and propose improvements in risk management and transparency. **Case Study:** Cambridge Analytica and Facebook	
	L8: Safety in AI: Mapping, measuring, and managing	**Classify and map AI safety risks in key domains:** Analyze and categorize safety risks, such as physical and legal harms, using the NIST AI RMF. Evaluate risks in real-world contexts, like healthcare and autonomous vehicles, identifying failure points and proposing safety measures. **Measure and monitor AI safety with risk assessment tools:** Apply safety metrics like adversarial robustness and model drift detection. Use monitoring techniques to assess real-time risks in dynamic systems, focusing on financial AI or autonomous driving. **Assess strategies to manage AI safety trade-offs:** Develop governance frameworks balancing safety, fairness, and innovation while engaging stakeholders. Critically evaluate trade-offs, like privacy vs. safety, using examples from legal AI and autonomous vehicles.	(Quiz-4)
5	L9: Robustness and Reliability in AI: Mapping. Measuring, and managing	**Apply and evaluate robustness techniques in machine learning models:** Implement and assess the effectiveness of techniques like adversarial training and regularization in mitigating distribution shifts and adversarial attacks. **Map and measure reliability risks using the NIST AI RMF:** Use the NIST AI Risk Management Framework to map and evaluate reliability risks, focusing on model performance, stability, and reproducibility across different scenarios. **Design and implement strategies to manage robustness and reliability risks:** Apply the NIST AI RMF to develop risk management strategies that improve model reliability, ensuring robust and stable performance in real-world applications.	

	L10: AI Governance	**Apply NIST AI RMF in real-world governance contexts:** Use the NIST AI RMF to assess governance challenges and propose risk management strategies, applying its core functions (Govern, Map, Measure, Manage) to mitigate risks in industry-specific AI systems. **Analyze global AI governance trends:** Compare and evaluate AI governance frameworks from the U.S., EU, and Singapore, using case studies to assess the effectiveness of these models in managing AI risks and ensuring compliance with current laws. **Assess ethical and social risks in AI systems:** Apply AI assurance methods, such as impact assessments and algorithmic audits, to identify and mitigate risks related to bias, transparency, and environmental harm in AI systems across various sectors.	(Quiz-5) (Individual Assignment 2)
6	Guest Lecture & Project Presentation	Guest Lecture: TBD and Team Project Presentation	(Quiz-6) Team Project Presentation