Carnegie Mellon University

94879 Fundamentals of Operationalizing AI

Meeting Days, Times, Location: MW – 9:30AM-10:50AM; HBH 2008

Semester/Year: Fall 2023

Units: 6**, Section(s):** A2

Instructor information

Name	Prof Anand Rao
Contact Info	anandr2@andrew.cmu.edu
Recitation	Friday – 2:00PM-03:20PM
Location	HBH 2008
Office location	Hamburg Hall 2105D
Office hours	Tuesday – 3:00PM-4:00PM (Also available through Zoom)

Course Description

Embark on a transformative journey with our course, "Fundamentals of Operationalizing AI: Mastering AI Lifecycle from Theory to Practice". This comprehensive program is meticulously designed to provide an in-depth understanding of the entire AI lifecycle. It covers a broad spectrum of topics, from business scoping, data management and engineering, to model development, deployment, and stewardship.

The course introduces the concept of Operationalizing AI (OAI), a critical aspect of AI implementation that is often overlooked. You will learn about its significance, the challenges it presents, its benefits, and the roles involved. We delve into related disciplines such as DevOps, DataOps, DevSecOps, MLOps, and AI Ops, focusing on the emerging best practices, roles, skills, capabilities, and governance across the AI lifecycle.

Upon completion, you will have a robust understanding of how to build and manage large-scale, production-quality AI systems that seamlessly integrate data, software, and models. Several guest lectures from industry practitioners and companies developing tools for operationalizing AI will give you a window into how organizations are operationalizing AI and creating the necessary tooling. This course is an invaluable resource for those aspiring to master the art and science of AI implementation, offering practical insights and knowledge that can be immediately applied in the real world. Join us and equip yourself with the skills to navigate the exciting world of AI.

Learning Objectives

1. Clearly explain the key components of the AI lifecycle, its maturity levels, and how they apply to various industries and functional areas.
2. Identify and describe the different stages of the AI lifecycle, including the specific artifacts, roles, skills, and capabilities required at each stage.

3. Create an end-to-end, top-down governance framework that ensures the consistent, efficient, and effective operation of AI systems.
4. Identify business problem spaces and match with appropriate technology solution spaces and apply techniques for validating models.
5. Assess the existing talent and skills within the organization and develop a plan to fill any identified gaps.
6. Develop a comprehensive strategy for operationalizing AI, including considerations for the development lifecycle, model deployment, monitoring, and change management.
7. Define and measure Key Performance Indicators (KPIs) for AI projects, calculate Return on Investment (ROI), and propose strategies for scaling and maintaining AI solutions.
8. Prepare and deliver a compelling presentation that communicates a comprehensive AI lifecycle strategy to an executive audience, highlighting the business value and potential impact of the proposed approach.

Learning Resources

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

1. Operationalizing AI: How to accelerate and scale across people, process, and platforms by John J. Thomas, William Roberts, and Paco Nathan, O'Reilly Media, March 2021.
2. Operating AI by Ulrika Jagare, Wiley, May 2022.
3. Designing Machine Learning Systems by Chip Huyen, O'Reilly Media, May 2022.

Assessments

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

Assessment	Percentage of Final Grade
Class Participation	10%
Individual Assignment	20%
Best 4 of 5 Quizzes	40%
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 Assignment:** Individual assignment will be based on answering the discussion questions in the industry-based synthetic scenarios that will be distributed at least one week before the assignment is due. See note on the use of generative AI tools in the Generative AI Guidance section.
- **Quizzes**: Five classroom/online quizzes will be administered during Week 2 to Week 6 (inclusive of both these weeks) of the course. Each quiz will be for 10 points 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 five quizzes without affecting the full score of 40% 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 Schedule

Date	Theme/Topic	Learning Outcomes Addressed	Assignments Due
M: Oct 23	Intro & Overview		
W: Oct 25	Models vs Software code (Synthetic Scenario)		
M: Oct 30	AI System Life Cycle	Clearly explain the key components of the AI lifecycle, its maturity levels, and how they apply to various industries and functional areas.	Quiz-1
W: Nov 01	Value & ROI estimation (Synthetic	Define and measure Key Performance Indicators (KPIs) for AI projects, calculate Return on	

	Scenario)	Investment (ROI)	
M: Nov 06	Model building and validation	Identify business problem spaces and match with appropriate technology solution spaces and apply techniques for validating models	Quiz-2
W: Nov 08	Model Deployment (Synthetic Scenario)	Analyze methods for deploying models at scale	
M: Nov 13	Model Monitoring	Evaluate challenges, need and techniques for monitoring data, models, and software	Quiz-3
W: Nov 15	Operationalizing AI Case Study - Guest Lecture (TBD)	Identify and solve challenges of applying operationalization in the industry.	
M: Nov 20	Trust Management	Create an end-to-end, top-down governance framework that ensures the consistent, efficient, and effective operation of AI systems.	
W: Nov 22	*No Class - Thanksgiving*		
M: Nov 27	People, Process, Organization	Assess the existing talent and skills within the organization and develop a plan to fill any identified gaps	Quiz-4
W: Nov 29	Guest Lecture (TBD)		Individual Assignment Due
M: Dec 04	Generative AI & LLM Ops	Assess current trends and the future of operationalizing AI	Quiz-5
W: Dec 06	Summary & Reflections	Recap of key learnings from course and final project presentation details	
TBD	Final Project Presentation		Final Team Presentation

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: October 23 (Monday): Introduction & Overview**
  - **Topics**
    - Introductions
    - Structure of the class
    - Expectations
    - What is “Operationalizing AI”?
    - Why study “Operationalizing AI”?
  - **Readings**:
    - [Thomas] Chapter 1: Current View and Challenges of AI Adoption
    - [Jagare] Chapter 1: Balancing the AI Investment
    - [Huyen] Chapter 1: Overview of Machine Learning Systems
- **Lecture 2: October 25 (Wednesday): Models vs Software Code**
  - **Topics**
    - Case Study Discussion: Bridging Two Worlds: The Trials and Triumphs of Integraing AI into Traditonal Sofware at FinSolutions
    - Differences between machine learning models and software code
    - Consequences of mistaking models for software
  - **Readings**:
    - [Huyen] Chapter 2: Introduction to Machine Learning Systems Design
  - **Additional Readings:**
    - *Data Scientists are from Mars and Software Developers are from Venus (Part 1). Anand Rao Towards Data Science. August 29, 2020.*
    - *Consequences of mistaking models for software (Part 2). Anand Rao. Towards Data Science. September 6, 2020.*
    - *Why are machine learning projects so hard to manage? Lukas Biewald, Medium, January 28, 2019*
    - *Create a common-sense baseline first. Rama Ramakrishnan, Medium. January 12, 2018.*
    - *A guide to different bias mitigation techniques in machine learning. Sourabh Mehta, Analytics India. April 2, 2022.*
- **Lecture 3: October 30 (Monday): AI System Lifecycle**
  - **Topics**
    - Software and Data Science Lifecycles
    - AI System Lifecyle
    - Introduction to MLFlow
  - **Readings**:
    - [Thomas] Chapter 4: Stages of AI Life Cycle
    - [Jagare] Chapter 3: Embracing MLOps
    -
  - **Additional Reading:**

        - *Model Evolution: From Standalone Models to Model Factory (Part 3)*,Towards Data Science, September 13, 2020.
        - *Model Lifecycle: From ideas to value.* Anand Rao, Towards Data Science, September 26, 2020.
        - *MLOps: Machine Learning Life Cycle*, ML4Devs
        - *MLOps: Definitions, tools and challenges.* Symeonidis G., et.al.,, arXiv:2201.00162v1, January 2022.
- **Lecture 4: November 01 (Wednesday): Value Scoping**
    - **Topics**
        - Case Study Discussion: ROI of AI: Navigating the Road to Value Realization at MidWest Financial
        - Developing a business case
        - ROI for AI
    - **Readings**:
        - [Jagare] Chapter 7: Achieving Business Value from AI
    - **Additional Reading:**
        - *Solving AI's ROI problem is not that easy.* Anand Rao, Tech Effect, July 20, 2021.
        - *How a Portfolio approach to AI helps your ROI*. Anand Rao, Tech Effect, September 9, 2021.
- **Lecture 5: November 06 (Monday): Value Discovery**
    - **Topics**
        - Model discovery steps
        - Data engineering
        - Feature engineering and feature stores
        - Model development
        - Model evaluation
    - **Readings**:
        - [Huyen] Chapter 6: Model Development and Offline Evaluation
    - **Additional Readings:**
        - [Jagare] Chapter 2: Data Engineering Focused on AI
        - [Huyen] Chapter 3: Data Engineering Fundamentals
        - [Huyen] Chapter 4: Training Data
        - [Huyen] Chapter 5: Feature Engineering
- **Lecture 6: November 08 (Wednesday): Value Delivery**
    - **Topics**
        - Myths of model deployment
        - Types of model deployment
        - Batch and Online architectures for inference
        - Deployment strategies
        - Model performance tuning
        - Docker
    - **Readings**:
        - [Jagare] Chapter 4: Deployment with AI Operations in Mind
        - [Huyen] Chapter 7: Model Deployment and Prediction Service
- **Lecture 7: November 13 (Monday): Value Stewardship**
    - **Topics**
        - Data drift and model drift
        - Continual learning

        - Infrastructure and tools for MLOps
    - **Readings**:
        - [Jagare] Chapter 5: Operating AI is different from Operating Software
        - [Huyen] Chapter 8: Data Distribution Shifts and Monitoring
        - [Huyen] Chapter 9: Continual Learning and Test in Production
        - [Huyen] Chapter 10: Infrastructure and Tooling for MLOps
    - **Additional Readings:**
        - *From concept drift to model degradation: An overview on performance-aware drift detectors. Bayram., et. Al., Knowledge- Based Systems 245, 2022.*
        -
- **Lecture 8: November 15 (Wednesday): Guest Lecture on Operationalizing AI**
    - **Topics**
        - Pharmacovigilance production Model & AI Industrialization at Pharma company
- **Lecture 9: November 20 (Monday): Trust Management**
    - **Topics**
        - NIST RAI Framework
        - Bias and Fairness
        - Explainability and Interpretability
        - AI Governance
    - **Readings**:
        - [Jagare] Chapter 6: AI is All About Trust
    - **Additional Readings:**
        - *Five Views of AI Risk: Understanding the darker side of AI (Towards Responsible AI — Part 1),* Anand Rao. Towards Data Science, November 28, 2020.
        - *Ten principles of Responsible AI for corporates* (Towards Responsible AI — Part 2), x Anand Rao. Towards Data Science, December 17, 2020.
        - Top-down and end-to-end governance for the responsible use of AI (Towards Responsible AI — Part 3). Anand Rao. Towards Data Science, January 19, 2021.
        - Six stage gates to a successful AI Governance (Towards Responsible AI – Part 4). Anand Rao. Towards Data Science, February 21, 2021.
- **Lecture 10: November 27 (Monday): People, Process, and Organization**
    - **Topics**
        - Emerging skills and capabilities
        - Operating models for AI
        - Center of Excellence
        - Human-centered AI
    - **Readings**:
        - [Thomas] Chapter 5: AI Center of Excellence
        - [Thomas] Chapter 2: Personas and Effective Communication Among Them
        - [Thomas] Chapter 3: Design Thinking
        - [Huyen] Chapter 11: The Human side of Machine Learning
- **Lecture 11: November 29 (Wednesday): Guest Lecture**
    - **Topics**
        - TBD
- **Lecture 12: December 4 (Monday): Generative AI and LLMOps**
    - **Topics**
        - Generative AI

        - Transformer Models – Large Language Models
        - Prompt Engineering
        - Model Adaptation
        - Model Evaluation
    - **Readings**:
        - An overview of Large Language Models (LLMs) by Mostafa Ibrahim, Weights and Biases, June 22, 2023.
        - Understanding LLMOps: Large Language Model Operations by Leonie, April 2023, Weights & Biases.
        - Essential guide to foundation models and large language models by Babar M. Bhatti, Medium, February 5, 2023.
        - Exploring opportunities in the Generative AI value chain by Tobias Härlin, Gardar Björnsson Rova, Alex Singla, Oleg Sokolov, and Alex Sukharevsky, McKinsey Digital, April 2023.
        - Emerging architectures for LLM Applications by Matt Bornstein and Rajko Radovanovic, Andreesen Horowitz, June 20, 2023.
    - **Additional Readings & Resources**:
        - Prompt Engineering Guide
        - Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., . . . Wen, J. (2023). A Survey of Large Language Models. ArXiv. /abs/2303.18223
        - What is LLMOps? Large Language Models' Ops, Architecture & Recommended tools by Arun, April 2023, Accubits.
        - A Developer's Guide To LLMOps: MLOps for Operationalizing LLMs by Hakan Tekgul and Aparna Dhinakaran, Arize.
        - Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., . . . Gui, T. (2023). The Rise and Potential of Large Language Model Based Agents: A Survey. *ArXiv*. /abs/2309.07864.
- **Lecture 13: December 6 (Wednesday): Conclusion and Recap**
    - **Topics**
        - Summary of topics
        - Key takeaways
        - Final Presentation