Carnegie Mellon University

Generative AI: Applications, Implications, and Governance - 94816

Location: HBH (Room 1005)

Semester: Fall 2023 (Mini 1)

Units: 6

Instructor information

Name	Jordan Usdan
Contact Info	Jusdan@andrew.cmu.edu
Office location	Virtual
Office hours	By appointment, please email

TA Information

TA name	Helom Berhane
TA Contact Info	Helom@cmu.edu
Office location	Virtual
Office hours	By appointment, please email

Course Description

We have entered the era of Generative AI, which holds transformational potential akin to the industrial revolution or the advent of the Internet. This technology is now capable of comprehending and generating language, code, images, and videos, performing routine digital tasks, and aiding humans in advancing scientific and creative fields.

These intelligent capabilities are powering novel AI user experiences such as OpenAI's ChatGPT and Dall-E, Google's Gemini, Microsoft's Copilot and various proprietary and open-source offerings. These innovations can boost human productivity and creativity, but their impact on society will be somewhat unpredictable long-term. By immersing ourselves in the study of Generative AI, we not only equip ourselves for personal growth but also become architects of a future where technology serves humanity's highest aspirations.

In this course, students will explore the broad impacts of generative AI by considering its applications, societal implications, and governance. We will explore how leveraging AI tools can help us as individuals thrive in our personal and professional lives. We will learn how both governments and private enterprises are seeking to develop beneficial and safe AI aligned with human needs.

This course will provide students with practical exposure to the latest AI technologies. Students will learn the art of prompt engineering, how to use AI as a research, writing and thinking tool, and strategies to avoid common AI failure points.

The first part of the course will cover how generative AI is created and applied, including the potential for it to augment human capabilities in beneficial ways. This includes a host of in-class activities using Microsoft Copilot, ChatGPT and other AI tools. This will include workshops on researching and writing with the help of AI.

The second part of the course will cover societal implications and governance, including public policy. Issues we will cover include:

- Economic and labor impact: Large generative models can be powerful tools to empower humans but could also automate many tasks that were previously done by humans, potentially leading to job loss and disruption in the job market.
- Information and ecosystem impact: Large generative models can be used to help moderate polarizing conversations, but they can also generate fake information (called hallucinations) and can be used to create deepfakes, poisoning the information commons.
- Data ownership impact: AI relies on training data in the form of text, images, and videos to learn patterns to understand and generate related output. However, open questions remain on how to compensate content owners and reward the creativity of content producers in the age of AI.

Class sessions and assignments will involve hands-on AI activities, including the creation of audio and video “deepfakes” and “red teaming” of models. We will host in-class discussion groups that dissect AI public policy debates across the US, EU and international realms. Additionally, we will explore techniques for responsible AI development and release at organizations.

**Note: AI tools such as ChatGPT and Microsoft Copilot were used to help generate themes to cover in this course and to find readings, generate activities to support this curriculum, and help edit this syllabus for clarity.*

Prerequisites

There are no prerequisites for this class. The course is designed for any graduate student to experience the latest AI technology and explore its societal implications. It is ideal for students on policy, business, design, data science or technology tracks.

Learning Objectives

Students will be able to:

1. Explore and utilize generative AI tools, learning hands-on ways to integrate them into daily routines through crafting text prompts for diverse tasks. Students will also learn strategies to avoid hallucinations, ensuring the generated outputs are accurate and reliable.

2. Discover the art of co-authoring with AI, understanding the advantages and limitations of language models in writing.

3. Gain a high-level understanding of the science behind generative AI, encompassing how it predicts text sequences and creates images. Students will grasp the fundamentals of AI training, fine-tuning, and inference while staying abreast of current trends in the field.

4. Delve into the societal ramifications of Generative AI through economic, political, ethical, and business perspectives. Engage in discussions to form opinions and debate the potential unforeseen effects of generative AI on society.

5. Devise public policy strategies aimed at alleviating the adverse impacts of this technology on society.

6. Pose insightful questions to experts and critically assess their viewpoints.

7. Learn the principles of responsible AI governance within organizations, including red teaming and responsible release strategies.

8. Articulate well-informed recommendations to decision-makers on complex issues, presenting concise and substantiated arguments.

Learning Resources

No textbooks are required. Readings will be available on Canvas.

Technology Resources:

- Recommended:
  - Required: Microsoft Copilot at copilot.Microsoft.com via your CMU login.

  - Recommended: also recommend you register for ChatGPT Plus, which is $20 per month. ChatGPT Plus: https://chat.openai.com/Plus
- Alternatives:
  - Poe.com – allows access to a variety of AI models
  - Perplexity.ai or labs.perplexity.ai allows access to a variety of AI models and real-time search + AI results.

Course Requirements

1. Students ***are required to bring a laptop or tablet with keyboard to class.***
2. Students are required to register for Copilot.Bing.com or subscribe for the $20 per month access to ChatGPT Plus during the length of the course.

Assessments

Assessment	Percentage of Final Grade
Class discussion participation grade	10%
Assignment #1 – Canvas Assignments	15%
Assignment #2 – Full AI Writing Exercise (in class)	10%
Assignment #3 – AI Productivity Log	10%
Assignment #4 – Red-teaming analysis (OR) Deepfake creation	35%
Assignment #5 - AI Policy Editorial (w/ AI Co-Author)	20%

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

Grade	Percentage Interval
A+	95-100%
A	90-94%
A-	88-90%

B+	85-87%
B	80-84%
B-	78-80%
C	70-78%
D	61-70%
R (F)	Below 60%

Grading Policies

- **Late-work policy**: *Late work will not be accepted unless the professor is notified before the deadline with an exceptional circumstance, exceptions will be permitted case-by-case*.
- **Attendance and/or participation policy**: In-person attendance is mandatory unless a valid reason is provided per CMU policies. Attendance will be taken by the TA.

Course Policies

- **Attendance & Participation**: Attendance and participation are a graded component of this course. I will be evaluating you based on contributions to class discussions, thoughtful commentary, and participation in Canvas discussions.
- **Academic Integrity & Collaboration**: This course follows all CMU rules on academic integrity.
- **Use of AI**: Use of AI is permitted as part of this course and is highly recommended as a tutor to help students understand concepts. Several assignments will explicitly incorporate the use of AI and others may disallow it. **Where it is not explicit, please cite your use of AI and do not pass off work substantially created by AI as your own**.
- **Accommodations for students with disabilities**: If you have a disability and require accommodations, please contact Catherine Getchell, Director of Disability Resources, 412-268-6121, getchell@cmu.edu. If you 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.
- **Statement on 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."

Course Schedule

Date	Class #	Topics	In Class Agenda	Assignments Due
8/28	1	**Generative AI Overview**, Copilot onboarding, prompt engineering techniques.	Course Overview Overview of Language Models Hallucinated class introductions In class exercise on LLM parameters Introduction to Prompt Engineering	**Pre-class survey**
9/4	2	**Technology behind Generative AI, Advanced AI productivity and authoring techniques,** exploring voice, image, and video generation	Gen AI Tech Overview What Can GPT-4 do now exercise Full AI In-Class Writing Exercise Advanced Writing and Research Workshop	**Canvas Assignment (In-Class)** **Assignment #2 (In-Class):** Writing by AI and self-evaluation.
9/11	3	**Generative AI Advanced writing techniques,** exploring advanced research, thinking and writing with AI **Deepfakes and Red Teaming overview, assignments provided**	Advanced AI Tech Demos (Voice, productivity, data) Overview of Risks and Benefits of AI In-class create your own Deepfake Discussion on Labor and the Economy Discussion on assignment #4	**Canvas Assignment (In-Class)**

			(Deepfake or Red Team)	
9/18	4	**AI Impact on the Information Ecosystem** **Mitigating disinformation**	AI Usage Log Discussion In class: How Do Your Politics Stack Up Against ChatGPT's? Conspiracy Theory Exercise Information Ecosystem Lecture Jailbreaking overview Red teaming exercises	**Assignment #3: AI Usage Log** **Canvas Assignment (In-Class)**
9/25	5	**AI governance and responsible release at companies**, framework for creating RAI products, (guest speaker) Discuss approaches to **deepfakes**, video capture sessions with Professor during break Class outing @ The Underground 8:45pm	AI Governance Lecture Guest Speaker AI Safety Class Outing	**Canvas Assignment (In-Class)**
10/2	6	**Accountability for Generative AI**, developing liability regimes, EU AI Act, White House voluntary commitments Discuss AI Policy Editorial	Code your own app with Claude Artifacts. Examples. Policy lecture Copyright Policy Discussion Discussion about assignment #5	**Canvas Assignment (In-Class)**

10/9	7	**Deepfakes and red-teaming assignments due, Pathways for global governance of AI**	Deepfakes + red-teaming presentations (select) Final lecture and wrap-up Post-class surveys	**Assignment #4** Red-teaming analysis (OR) Deepfake creation (Due 24 hours prior to class) **Assignment #5 AI Policy Editorial (Due before Mini 1 finals deadline – TBD)**

Reading assignments due before each class:

Class #1 Aug 28: Generative AI Overview, OpenAI onboarding, prompt engineering techniques

The first class will cover the historical backdrop of past technological revolution and societal consequences for them, setting up the high stakes of steering AI technology to benefit humanity. After an introduction of how Large Language Models work, we will start using several AI tools to see how they can be applied in practice and learn prompt engineering techniques to get the models to provide helpful outputs. We will also consider how to avoid common failure modes.

Generative AI overview – read prior to first class

- YouTube - Generative AI in a Nutshell - how to survive and thrive in the age of AI (youtube.com)
- Metz, Cade. "Microsoft Says New A.I. Shows Signs of Human Reasoning." *The New York Times*, 17 May 2023.
- Grace, Katja, et al. "Thousands of AI Authors on the Future of AI." Preprint, January 2024. (*Skim charts and/or use AI to summarize*)
- Peters, Jay. "AI is confusing — here's your cheat sheet - The Verge." Accessed 24 Jul 2024.

The art of prompt engineering – reference materials

- Mollick, Ethan, "An Opinionated Guide to Which AI to Use: ChatGPT Anniversary Edition (oneusefulthing.org)." 7 Dec 2023.
- Alston, Elena. "What are AI hallucinations—and how do you prevent them?" *Zapier*, 5 Apr. 2023
- OpenAI. "OpenAI Prompt Cookbook." *GitHub*, n.d.. Accessed 21 Aug. 2023.
- Amattrian, Xavier, Prompt Engineering and Design: Introduction to Advanced Methods, arXiv, May 2024

Class #2 Sep 4: Technology behind Generative AI, Advanced AI writing techniques, exploring voice, image, and video generation

The second class will start with a deeper dive into how Large Language Models and image creation models are developed, including the data inputs and the stages of training and development. We will then cover how these base models can be integrated into more sophisticated software applications and learn how to use several of these more advanced applications. Finally, we will see demonstrations of the most cutting-edge applications using Generative AI and begin to imagine how these technologies might impact society.

The technology behind Generative AI

- Fenjiro, Youssef. "ChatGPT & GPT 4, How it works ?. What is ChatGPT & GPT4?." Medium, 14 Apr. 2023.
- Optional readings:
  - Bowman, Samuel R. "Eight Things to Know about Large Language Models." Courant Institute of Mathematical Sciences, New York University, n.d.
  - Wolfe, Cameron, Explaining ChatGPT to Anyone in <20 Minutes (substack.com)

Future paths:

- Wang, Sarah and Xu, Shangda. "The Next Token of Progress: 4 Unlocks on the Generative AI Horizon." Andreessen Horowitz Blog, 23 Jun. 2023.
- ChatGPT 5 and Beyond: OpenAI's Five-Level Roadmap to AGI Unveiled | by Antonello Sale

| Jul, 2024 | Medium

Advanced AI uses and techniques

- What's something you use ChatGPT for that you're sure no one else does? : r/ChatGPT (reddit.com)
- Mollick, Ethan. "How to Use AI to Do Stuff: An Opinionated Guide." One Useful Thing, n.d., . Accessed 21 Aug. 2023.

Class #3 Sep 11: Advanced productivity techniques; Risks and benefits of Generative AI overview, impact on jobs and the economy

The third class will provide a holistic perspective of the societal impacts of Generative AI. We will consider how AI can be both a powerful tool and a weapon in many dimensions. We consider the impacts of generative AI to start to develop a point of view on what benefits and risks are probable and impactful versus others that might be unlikely and low impact.

Risks of Generative AI

- Weidinger, Laura et al. "Taxonomy of Risks posed by Language Models." FAccT '22, 2022
- YouGov. "What Americans think about ChatGPT and AI-generated text." YouGov, 5 Aug. 2023
- Optional readings:
    - "A Hazard Analysis Framework for Code Synthesis Large Language Models." FAccT '22, 2022, [1]
    - "Ethical and social risks of harm from Language Models." arXiv preprint arXiv:2112.04359, 2021,
    - EPIC. "Generating Harms: Generative AI's Impact & Paths Forward." EPIC, 2023, [4].
    - "AI scam artists impersonate familiar voices to scam the rest of us." CBS Pittsburgh, 11 August 2023

Generative AI Benefits

- Andreessen, Marc. "Why AI Will Save the World." Andreessen Horowitz Blog, 6 Jun. 2023.
- "The Economic Potential of Generative AI: The Next Productivity Frontier." McKinsey & Companyee, May 2023. (*Skim, view charts in Section 1 and 3*)d

Labor and the Economy Discussion Pieces:

- Levin, Blair and Downes, Larry, Is AI Really a Job-Killer? A Little Yes and a Big No. LinkedIn June 2024 .
- Kelly, Jack. Goldman Sachs Predicts 300 Million Jobs Will Be Lost Or Degraded By Artificial Intelligence (forbes.com). Forbes. 31 March 2023
- Baily, Martin Neil, et al. "Machines of Mind: The Case for an AI-Powered Productivity Boom." Brookings, 10 May 2023.
- Brynjolfsson, Erik. "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence." FAccT '22, 2022, 1[1]
    - *This is a very long read, so please have a "chat" with this PDF using AI by uploading this to ChatGPT (e.g.) example questions: What is the Turing Trap? Why are humans likely to fall into the Turing Trap? What are the problems caused by falling into the Turing trap? What is the relevancy of the Turing Trap to Generative AI? How can society avoid the Turing Trap? What are criticisms of the Turing Trap concept?*

Introduction to red teaming:

- Oremus, Will. "Meet the hackers who are trying to make AI go rogue." The Washington Post, 8 Aug. 2023.
- OpenAI, "Preparedness Challenge (openai.com)", Last Accessed 28 Oct. 2023.

Class #4 Sep 18: Generative AI impact on information ecosystem, mitigating disinformation (guest speaker)

The fourth class will take a deep look at the information ecosystem impact of Generative AI, including the news ecosystem. We will also consider the potential impacts on cybersecurity and consumer fraud, including via voice AI clones and other deepfakes. We will have a guest speaker who will help us consider impacts and policy options for AI and disinformation.

Reading prior to first exercise

- Study on AI Chatbots Reducing Conspiracy Beliefs (suchscience.net)

Information ecosystem impact

- Sadeghi, McKenzie, et al. "Tracking AI-enabled Misinformation: Over 400 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools." NewsGuard, 7 Aug. 2023.
- Vincent, James. "How Will Artificial Intelligence Change Journalism?" New York Magazine, 21 Aug. 2023.

Impact on Disinformation

- Thomspon, Stewart, How 'Deepfake Elon Musk' Became the Internet's Biggest Scammer. New York Times. Aug 14, 2024
- Kapoor, Sayash and Narayanan, Arvind. "How to Prepare for the Deluge of Generative AI on Social Media." Knight First Amendment Institute, 13 Apr. 2023.
- Kinsella, Brett. "TikTok's New Rules on Deepfakes and Other Synthetic Media." Substack, 23 Mar. 2023.
- Wolf, Wolf! Alarm Over Disinformation and The Liar's Dividend | CCCB LAB. CCCB LAB, 2023. [1]

Class #5 Sep 25: AI governance and responsible release at companies, framework for creating RAI products, (guest speaker), Live Red Teaming exercise

The fifth class will study how AI models and products are being developed and released at companies via the implementation of Responsible AI development processes. These processes include harms analysis, red-teaming, and responsible release approaches. We will hear from a guest speaker from industry on Responsible AI and red-teaming and conduct our own read-teaming exercise.

AI Deployment Risks to Organizations

- "Our thinking: The flip side of generative AI." KPMG, 6 July 2023.

AI Release Methods to Mitigate Harms

- Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." arXiv:2302.04844 [cs.CY], 5 Feb. 2023.
- "Lessons learned on language model safety and misuse." OpenAI, 3 March 2022

Responsible AI Product Development and Red Teaming

- Crampton, Natasha. "Microsoft's framework for building AI systems responsibly." Microsoft On the Issues, 21 June 2022, 5. See also "Responsible AI Standard." Microsoft, June 2022 (and Skim link to Responsible AI Standard)
- Red team approach Crescendo (crescendo-the-multiturn-jailbreak.github.io)
- (Optional) Schuett, Jonas, Anka Reuel, and Alexis Carlier. "How to design an AI ethics board." arXiv:2304.07249 [cs.CY], 14 Apr. 2023

Being human in the Age of AI:

- Brooks, David "Opinion | In the Age of A.I., Major in Being Human - The New York Times (nytimes.com)." New York Times, 2 Feb 2023

Class #6 Oct 2: Accountability for Generative AI, developing liability regimes, EU AI Act, White House voluntary commitments, Open Source models

The sixth class will consider policy options to maximize the potential and minimize the harms of Generative AI. We will analyze and compare EUs AI Act, the US Senate's approach, and corporate perspectives. We will also have a discussion on copyright.

Copyright discussion reading:

- Soffier, Ariel, Copyright Fair Use Regulatory Approaches in AI Content Generation (techpolicy.press); 8 August 2023.
- Evans, Bendict "Generative AI and Intellectual Property", 27 August 2023

Policy Principles

- "Blueprint for an AI Bill of Rights | OSTP | The White House." The White House, 4 Oct. 2022,
- Smith, Brad. "How Do We Best Govern AI?" Microsoft On the Issues, 25 May 2023,

Comparative Policy Approaches

- "A Law for Foundation Models: The EU AI Act Can Improve Competition and Innovation." OECD.AI,
- Engler, Alex. Proposing the CASC: A Comprehensive and Distributed Approach to AI Regulation (techpolicy.press). Tech Policy Press. 23 Aug 2023
- Matthews, Dylan. "The AI Rules That US Policymakers Are Considering, Explained." Vox, 1 Aug. 2023,

Perspectives on EU AI Act

- Optional readings:
  - GitHub et al. "Supporting Open Source and Open Science in the EU AI Act." GitHub Blog, 26 July 2023, [8].

Accountability for AI Agents (supplementary reading):

- Practices for Governing Agentic AI Systems, OpenAI, December 2023

Class #7 Oct 9: Pathways for global governance of AI, *approaches to democratic input into AI development (may change based on events, interest/engagement on topics)*

- *TBD – based on current events and class discussions.*
- Optional reading: www.situational-awareness.ai