Case Studies in Public & Private Policy Challenges of Artificial Intelligence
Harvard Law School
Harvard Kennedy School of Government
Harvard John A. Paulson School of Engineering & Applied Sciences
Spring 2024

Instructor: Jonathan Zittrain
Teaching Assistants:
Kalie Mayberry | Will Marks | Tom Zick

Monday & Tuesday: 3:45pm - 5:15pm

Location: WCC 3019

COURSE OVERVIEW

The use of machine learning has skyrocketed in recent years, becoming embedded retail and wholesale across society without substantial reflection on its implications. Through engagement with those building some of the most provocative models and tools – many of which have become part of the public imagination – we will see what gives their builders pause; reflect on possible solutions or mitigations; and develop suggestions about what they might be missing in their own canvassing of the ethical and policy terrain.

Please contact Will Marks <wmarks@jd25.law.harvard.edu> with any questions relating to course content or logistics.

COURSE MATERIALS

We've planned for a manageable amount of readings for this course. As discussion will be launched from the assigned readings, we require that students prepare carefully and completely for each session. In a discussion-based seminar format, failure to complete the readings will likely be both obvious and an impediment to the flow of class.

Most course readings are available on this course's **H2O playlist here**. Links to readings are also provided in the reading list that follows. Additional materials may be circulated via email or Canvas in advance of class.

ASSIGNMENTS

Students will be assessed on the basis of participation (including attendance), in-class activities (25%), as well as regular online engagements and written assignments (75%).

COURSE MATERIALS

We have worked to ensure that the readings for this course are very manageable. As discussion will be tightly focused around assigned readings, we require that students prepare carefully and completely for each session. Given our discussion-based seminar format, failure to complete the readings will be both obvious and an impediment to the flow of class. All course readings

are available on this course's H2O playlist here. Links to readings are also provided in the reading list that follows.

ACADEMIC HONESTY

Please refer to the Harvard Law School Handbook entry on Academic Honesty. And if in doubt, ask!

ATTENDANCE POLICY

Attendance is required for all class sessions. Attendance will be taken for every session, and unexcused absences will count against a student's participation grade. If you must miss a class, please reach out to the teaching team to make them aware you will not be attending class and receive instruction for writing a short reflection paper if they approve your absence as excused. You may only use the reflection paper to make up two classes: more than two unexcused absences will result in a decrease in grade for the course. Please see the HLS GPT Policy for more guidance.

ACCESSIBILITY POLICY

Harvard Law School is dedicated to facilitating equal access for students with disabilities and to cultivating a campus culture that is sensitive and responsive to the needs of students. To request an accommodation for a disability during the course, students are welcome to reach out to Accessibility Services at accessibility@law.harvard.edu or at 617-495-8773. Additional information, including how to register for accommodations, can be found on the HLS Accessibility Services Resources Page.

CHATHAM HOUSE PRACTICE

Course sessions are not to be recorded except by HLS according to HLS guidelines. What's said in class can be shared only without attribution, unless permission from a speaker is given otherwise.

READINGS AND SESSION INFORMATION

Week 1: Setting the Stage – Intro to Machine Learning & AI (01/22 & 01/23)

What is AI in 2024, and what consensus is there about its capabilities and trajectory?

- **Assignment:** On-boarding to BKC Project Loom:
    1. Create a Profile
    2. Post short introduction to "Introductions" Channel by Sunday, 1/28 at 11:59pm
    3. Post questions you have from readings in "Ask" Channel, especially if you are less familiar with machine learning!
- **Readings:**
    - Monday **(If you are less familiar with machine learning, start here!)**
        - "Visualizing the deep learning revolution" by Richard Ngo, 2023
            - *This reading shows, by examples across a variety of application areas, the dramatic increase in capability of AI systems over the past few years. How well does the essay make the case that these improvements can and will continue at the current pace? How much of our senses of societal*

problems and opportunities, and policy interventions, depend on accurately predicting AI's capabilities in the next interval?

- "Machine Learning: A Primer" by Lizzie Turner, 2018
  - *This reading offers a general view of machine learning and serves as background for ML that we'll build on over the semester.*
- "What Is ChatGPT Doing … and Why Does It Work?—Stephen Wolfram Writings" by Stephen Wolfram, 2023
  - *This reading is important to begin to understand how LLMs and ChatGPT work. The LLM underlying ChatGPT is simply generating text by repeatedly predicting the next token, given its training on lots of tokens. What types of text do you think would be challenging to generate well in this way?*
- The Building Blocks of Interpretability
  - *This reading is important to begin to understand how neural networks represent and combine concepts. What kind of interfaces into a neural network model would help you better understand how the model is operating? What kind of interfaces would help to increase your trust in how the model is operating?*

- Tuesday
  - **Primary**
    - *A Path Toward Autonomous Machine Learning* by Yann LeCun, 2022 (read pg 1-9)
      - *This reading is a leading AI researcher's perspective on creating autonomous intelligent agents. To what degree do you think autonomous intelligence agents must go through a similar development trajectory as humans (e.g., Figure 1) to achieve the level of broad competency humans have? To what degree do you think having a biologically-inspired architecture (e.g., Figure 2) is important to developing autonomous intelligent agents? Does this paper, from the portions you've read … make sense?*
    - "The Bitter Lesson" by Rich Sutton, 2019
      - *This reading is a leading AI researcher's perspective from 2019 on how the field of AI research oversteers toward "building in how we think we think we think" rather than relying on scaling computation. Can you think of counter examples where building in how we think that we think won't inhibit progress in the long run? What does the future of academic AI research look like if large, costly computational resources are required to make progress?*
  - **Recommended**
    - *The Mythos of Model Interpretability* by Zachary Lipton, 2017
      - *This paper disambiguates various notions of interpretability. For LLMs, what definition of interpretability is most relevant? How does this vary by application area? For image generation models, what definition of interpretability is most relevant? How does this vary by application area?*

Week 1.5: Machine Learning More in Depth (Friday 01/26, 12 - 2pm via Zoom)

- **Discussion leader:** Josh Joseph
- This session was recorded and recording posted on Canvas & Project Loom!
- Here are Google Slides for reference.

Week 2: Setting the Stage – AI Ecosystem of 2024 (01/29 & 01/30)

How is AI developed and by whom? How should it be developed and who can/should decide that? How should decisions be made about development?

- **Guest:** Pablo Arrendondo (CaseText)
    - *"Pablo Arredondo is the Co-Founder and Chief Legal Research Officer at Casetext. Casetext is a legal research platform where primary materials are linked to secondary analysis. Pablo is also a CodeX fellow at the Stanford Center for Legal Informatics where his work focuses on civil litigation in common law jurisdictions, with an emphasis on how litigators access and assemble the law. Prior to joining Casetext, he founded the legal research technology startup Occam and has represented leading technology companies in patent litigation. Before attending law school, he worked at the Center for Bioethics at the University of Pennsylvania School of Medicine."*
- **Assignment:** From Week 2 readings, bring 1-3 questions to your group. Within your groups during class on ~~Monday~~ Tuesday, discuss which 1 question you want to tackle as a group. Draft a post in "Class Only: Chat & Questions" with a copy of your question, the week it applies to, and your responses to the question. Group posts are due by ~~Tuesday, 1/30~~ Wednesday, 1/31 at 11:59pm. Before class on Monday, 2/05, read and comment on another group's response.
- **Readings:**
    - Monday
        - *Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries* by Robert Gorwa & Michael Veale, 2024 (S.2 and S.4 + skim at least one case study)
            - *This reading begins to explore the emerging business models around large language models and examines the implications of technical challenges on governance efforts and existing law. Do you agree with how the authors delineated between parts of the LLM "stack"? If so, does this change how authors should interpret the law? Can you think of other case studies that would test the analysis explored here?*
        - "How OpenAI is boosting scrutiny of Microsoft's market power" by Rebecca Klar, 2023 [Additionally on Canvas]
            - *This reading will give you an idea of the current toolkit for governing markets as applied to OpenAI and Microsoft. Where do you think the power lies in OpenAI's corporate governance structure? Are there changes the parties could make that would further protect them from undesired scrutiny? The structure of the emerging LLM industry may make certain*

governance solutions untenable while boosting the attractiveness of others – should we tailor the solutions?

- Tuesday
  - FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI | The White House
    - *This is an example of government-fostered "self-regulation." What might the motivations for such an approach be, as compared to other models? Are the implicated companies what you would call "AI companies"? Does it matter?*
  - FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence | The White House
    - *This may be an example of a "collaborative governance" approach as it involves a combination of industry disclosure/standard building as well as government direction. In what way does it differ from the voluntary commitments in theory or practice?*
  - "France's Mistral takes a victory lap" by Derek Robertson, 2023
    - *This article explores the politics and stakes of the most recent AI Act vote. What are some of the apparent tensions faced by regulators in crafting AI legislation? Are there ways to reconcile them?*
  - "European Union squares the circle on the world's first AI rulebook" by Luca Bertuzzi, 2023
    - *This article provides an explainer of the AI Act. The AI Act will be published in the Official Journal of the EU and will enter into force 20 days following publication. Provisions related to prohibited AI systems are set to become enforceable six months after the Act is finalized and provisions related to General Purpose AI will become enforceable 12 months after this date. The rest of the AI Act is expected to become enforceable in 2026, but this is all still provisional at this stage. The article also discusses what the current exemptions are.*
  - "European Commission welcomes political agreement on Artificial Intelligence Act" by European Commission, 2023
    - *This article is the official Press Release from the European Commission regarding the AI Act. Who are the regulated parties? How does this affect the AI ecosystem?*
  - "European Commission AI Act Q&A" by European Commission, 2023
    - *This article is a helpful FAQ regarding the European AI Act.*

Week 3: Setting the Stage – The Regulator's Toolbox (02/05 & 02/06)

How do governance levers get pulled? What is the state of AI in United States Law? Global Law?

- **Cases:**
  - Sony Corporation of America v. Universal City Studios, 1984
  - Metro-Goldwyn-Mayer Studios Inc v. Grokster, LTD, 2005
  - NY Times v. OpenAI, 2023
- **Assignment:** No assignment this week - please note paper assignment due Week 5

- **Readings:**
  - Monday
    - "The Law of the Horse: What Cyberlaw Might Teach" by Lawrence Lessig, 1999 (Read 501-505; 546)
      - *The reading provides a classic introduction to cyberlaw's relationship to traditional law, demonstrating the forces of code, norms, markets, and law.*
    - *Sony Corporation of America v. Universal City Studios, 1984 [shortened version on Canvas]*
      - *This is a seminal case when it comes to the response of copyright to new technologies. Which legal theories does the court endorse or reject? Is this reasoning aligned with the outcome? Are there non-legal theories the court may be implicitly considering, and if so, what are they?*
  - Tuesday
    - Choose two of three:
      - *Metro-Goldwyn-Mayer Studios Inc. V. Grokster, Ltd., 2005 [shortened version on Canvas]*
        - *We revisit what happens when new technology erodes long-held notions of property. Compare the reasoning to that entertained in Sony – where do they overlap? Why do you think the outcomes are drastically different? (Are they?)*
      - "Provable Copyright Protection for Generative Models – Windows On Theory" by Boaz Barak, 2023
        - *This article explores a technical approach to mitigating potential copyright concerns. Are they the correct ones? Can you think of other technical approaches? Who might be happy with this, and who not? Are there any inherent limitations to the role we can expect technical solutions to play?*
      - *NYTimes v. OpenAI, 2023 [focus on paragraphs 1-9, 75 - 81, 98 - 100, 124 - 126, 169 - 180 + skim the rest for anything that jumps out at you!]*
        - *This is the complaint that NYTimes filed against open AI. What are the most compelling sections? Is there an ethical underpinning, or does it appear purely doctrinal?*

Week 4: Setting the Stage – Normative Frameworks & Human Thriving (02/12 & 02/13)

Why are we passing laws? And how to argue with humans that might not agree with you?

- **Case:** Normative Frameworks
  - **Guest:** Terry Fisher (2/12)
    - *"Professor Fisher received his undergraduate degree (in American Studies) from Amherst College and his graduate degrees (J.D. and Ph.D. in the History of American Civilization) from Harvard University. Between 1982 and 1984, he served as a law clerk to Judge Harry T. Edwards of the United States Court of Appeals for the D.C. Circuit and then to Justice Thurgood*

Marshall of the United States Supreme Court. Since 1984, he has taught at Harvard Law School, where he is currently the Wilmer Hale Professor of Intellectual Property Law. His academic honors include a Danforth Postbaccalaureate Fellowship (1978-1982) and a Postdoctoral Fellowship at the Center for Advanced Study in the Behavioral Sciences in Stanford, California (1992-1993)."

- **Assignment**:
  - 3-5 Page Single Spaced Paper Assignment #1 due next Monday, 2/19 by noon [submit on Canvas]:
    An AI case studies classmate suggests a business opportunity that might also have a beneficial social impact, particularly for vulnerable populations. Specifically, ongoing incidents of both suicide and domestic violence might be reduced by some combination of (1) modeling who might be at risk of suicide at a phase where supportive interventions could be helpful; and (2) AI-facilitated monitoring of rooms containing microphones on smart phones and digital assistants could automatically discover when an incident of domestic violence is likely in progress. No recordings would be made or shared unless an overwhelmingly high likelihood of violence-in-progress were found.

    Prepare an analysis of these opportunities as you consider your potential interest in the opportunity, specifically weighing in on at least one of the following aspects, for at least one of the two ideas above (it need not be both), including but not limited to: (1) are these worthy ideas in theory?; (2) what would it look like to attempt to implement them within the current ecosystem of relevant companies and service providers, whether through private or public policy; (3) if one or both were to move forward, what sorts of safeguards could be put in place to minimize chances of abuse or of misidentification of harmless activity?

- **Readings**:
  - *The Moral Foundations of Politics* by Ian Shapiro, 2003 (Read 18-25) [On Canvas]
    - *What is classical utilitarianism? Do we believe humans seek to maximize pleasure and minimize pain? What would it mean to set up society to do the same?*
  - *Anarchy, State, and Utopia* by Robert Nozick, 1974 (Read 41-45). [On Canvas].
    - *Should we cater to the wants and needs of those who extract more utility than do others? Should we value more than the maximization of pleasure/minimization of suffering?*
  - "The Ones Who Walk Away From Omelas" by Ursula K LeGuin, 1973
    - *This narrative questions what sacrifices the majority should be willing to impose on the few for the good of the many. Would you walk away from Omelas?*
  - *Famine, Affluence & Morality* by Peter Singer, 1972 (Read 229-235) [Additionally on Canvas]
    - *This article argues "if it is in our power to prevent something very bad from happening, without thereby sacrificing anything else morally significant, we ought,*

morally, to do it." If so, can we justify the decisions we make considering the poverty, famine, and death that exists in the world? What sacrifices are the fortunate morally obligated to make for the few?

- "Effective Altruism Committed the Sin It Was Supposed to Correct" by Annie Lowrey, 2022 [Additionally on Canvas]
  - *What is effective altruism? To what extent was SBF's downfall reflective of a flaw with the individual? With the movement? With the underlying ideology?*
- Deontological Ethics, The Stanford Encyclopedia of Philosophy, 2007
  - *What is deontology? What is consequentialism? Does process matter, or outcome? When designing systems where outcomes are not givens, what should we take into account?*
- *The Problematics of Moral and Legal Theory*, by Richard Posner, 1997 (Read sections A, F-I) [Additionally on Canvas]
  - *What is "academic moralism," and does moral theory provide "a solid basis for" moral decision making? Do you agree with Posner? What do we think about formalized ethics?*

Week 5: Emotional AI (02/19 & 02/20)

What ethical considerations should be taken into account when designing emotionally intelligent systems?

- **Case:** Affectiva
  - **Guest:** Ben Reis (02/19)
    - *"Ben Reis is Director of the Predictive Medicine Group and a member of the Faculty of Harvard Medical School and the Boston Children's Hospital Computational Health Informatics Program. His research focuses on understanding the fundamental patterns of human disease and on developing novel approaches for predicting disease. He has conducted large-scale population studies of COVID-19 vaccine effectiveness and safety, and developed novel methods for tracking and understanding pandemics through digital information sources. He has created systems that allow doctors to predict dangerous clinical conditions years in advance, including suicide and domestic abuse, as well as predictive pharmacology systems able to identify life-threatening adverse drug side-effects years in advance. He has advised the US government on establishing national biodefense systems, the Hong Kong government on building health infrastructure in response to pandemics, and various governments on establishing biodefense systems in advance of hosting the Olympic Games. He has been honored at the White House for his work on harnessing social networks to promote health, and was named one of the top health innovators in the world by the US State Department, USAID and NASA."*
  - **Guest:** Rana el Kaliouby (02/20)
    - *"Rana's life work is about humanizing technology before it dehumanizes us. She is an Egyptian-American scientist, entrepreneur, angel investor, author,*

and an AI thought leader on a mission to bring emotional intelligence to our digital world. She is the Deputy CEO at Smart Eye and formerly, Co-Founder and CEO of Affectiva, an MIT spin-off and category defining AI company. Rana realized a successful exit for Affectiva in June 2021 when the company was acquired by Smart Eye, where she is currently focused on scaling the company to a global AI powerhouse. She is also an executive fellow at the Harvard Business School where she teaches about AI and startups. Her bestselling memoir, Girl Decoded: A Scientist's Quest to Reclaim Our Humanity by Bringing Emotional Intelligence to Technology (Penguin Random House, April 2020), follows her personal journey, growing up in the Middle East and moving to the United States to become an entrepreneur, juxtaposed against her work building Emotion AI."

- **Assignment:**
    - 3-5 Page Single Spaced Paper Assignment #1 due next Monday, 2/19 by noon [submit on Canvas]:
      An AI case studies classmate suggests a business opportunity that might also have a beneficial social impact, particularly for vulnerable populations. Specifically, ongoing incidents of both suicide and domestic violence might be reduced by some combination of (1) modeling who might be at risk of suicide at a phase where supportive interventions could be helpful; and (2) AI-facilitated monitoring of rooms containing microphones on smart phones and digital assistants could automatically discover when an incident of domestic violence is likely in progress. No recordings would be made or shared unless an overwhelmingly high likelihood of violence-in-progress were found.

      Prepare an analysis of these opportunities as you consider your potential interest in the opportunity, specifically weighing in on at least one of the following aspects, for at least one of the two ideas above (it need not be both), including but not limited to: (1) are these worthy ideas in theory?; (2) what would it look like to attempt to implement them within the current ecosystem of relevant companies and service providers, whether through private or public policy; (3) if one or both were to move forward, what sorts of safeguards could be put in place to minimize chances of abuse or of misidentification of harmless activity?

    - **GROUP 1:** Firestarter Group Assignment + Individual Posts due Wednesday, 2/21 by 11:59pm. All other students expected to engage in discussions.

- **Readings:**
    - Monday
        - *Validation of an Electronic Health Record–Based Suicide Risk Prediction Modeling Approach Across Multiple Health Care Systems*, by Yuval Barack-Corren et al, 2020 [PDF on Canvas]

            - *This article uses electronic health records to predict incident suicide attempts. How do they situate their analysis within a social context? How does this compare to the other Ben Reis piece from 2009?*
        - *Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study* by Ben Reis, Isaac Kohane & Kenneth Mandl, 2009
            - *This article uses patients' historical records in an attempt to predict a patient's future risk of receiving a diagnosis of domestic abuse. Where do you think such predictions can fit in a social context? How do the authors grapple with this?*
    - Tuesday
        - "Emotion AI, explained" by Meredith Somers, 2019
            - *This article explains "emotion AI" and introduces us to Affectiva. Which industries are using emotion AI? Are there certain industries you feel would benefit greatly from using the tech? Certain ones you feel should refrain from using emotion AI?*
        - "Time to regulate AI that interprets human emotions" by Kate Crawford, 2021 [Additionally on Canvas]
            - *This article calls for the regulation of emotion AI. Does emotion AI work? If you are worried about the risk the technology poses, are you more worried about emotional AI working well or working poorly?*
        - Further Recommended:
            - "Smile if you think robots can read our emotions" by Lisa Feldman Barrett, 2017 [Additionally on Canvas]
                - *This article contends that universal emotions do not exist. If they do not, is the mission to build emotion AI destined to fail?*

Week 6: Bias & Automation in the Judicial System (02/26 & 02/27)

What happens when we bring predictive elements into the judicial system?

- **Case:** Clearview AI
    - **Guest:** Hoan Ton-That (02/26)
        - *"Hoan Ton-That is the CEO and co-founder of Clearview AI, which is based in New York City and has created the next generation of facial recognition technology. Clearview AI's bias-free algorithm can accurately find any face out of forty billion images it has collected from the public internet. It is used by law enforcement to solve crimes, including financial fraud, human trafficking, and crimes against children. A self-taught engineer, Mr. Ton-That is of Vietnamese and Australian heritage. His father's family was descended from the Royal Family of Vietnam. As a student, Mr. Ton-That was ranked #1 solo competitor in Australia's Informatics Olympiad. He was ranked #2 guitarist under age 16 in Australia's National Eisteddfod Music Competition. At the age of 19, Mr. Ton-That moved from Australia to San Francisco to focus on his career in technology. He created over twenty iPhone and Facebook applications with over 10 million installations, some of which ranked in the App Store's Top 10. Mr. Ton-That moved to New York City in*

2016. In 2017, Mr. Ton-That co-founded Clearview AI, where he developed the technology, raised capital, and built the team and product."

- **Guest:** Sandra Mayson (02/27)
  - *"Sandra Mayson is a Professor of Law at University of Pennsylvania's Carey Law School. Mayson researches and writes in the fields of criminal law, constitutional law, and legal theory, with a focus on the role of preventive restraint in the criminal legal system. Her academic work draws on her experience as a trial lawyer at Orleans Public Defenders, where she represented indigent clients in criminal proceedings and trained public defenders on immigration-sensitive defense practice. Mayson clerked for Judge Dolores K. Sloviter on the U.S. Third Circuit and Judge L. Felipe Restrepo in the Eastern District of Pennsylvania. Prior to joining the Penn faculty she was on the faculty of the University of Georgia School of Law, where she received the C. Ronald Ellington Award for Excellence in Teaching in 2020."*

- **Assignment:**
  - **GROUP 7:** Firestarter Group Assignment + Individual Posts due Wednesday, 2/28 by 11:59pm. All other students expected to engage in discussions.

- **Readings:**
  - Monday
    - *The following pieces shed light on the business of Clearview AI, a facial recognition company. Are you okay with the way the company trained its model? If you have any reservations about the use of the technology, why do you have them? Are you worried that the technology is too good at identifying individuals or that it is not good enough at doing so? Is it possible to be concerned about both?*
      - "The facial-recognition app Clearview sees a spike in use after Capital attack", by Kashmir Hill, 2021
      - "Facial Recognition Goes to War", by Kashmir Hill, 2022
      - "Clearview's Facial Recognition App Is Identifying Child Victims of Abuse", by Kashmir Hill & Gabriel J.X. Dance, 2020
      - "Clearview AI settles suit and agrees to limit sales of facial recognition database", by Ryan Mac & Kashmir Hill, 2020
      - "Long-running Clearview AI class action biometric privacy case settles" by Suzanne Smalley, 2023
      - "Clearview AI used nearly 1m times by US Police" by James Clayton & Ben Derico
  - Tuesday
    - *Governing by Algorithm? No Noise and (Potentially) Less Bias* by Cass Sunstein, 2022 (Read §§ I & II)

        - *This article presents an opinion that is heterodox in the legal academy but quietly fairly prevalent in many governments. What is Sunstein's main argument about the efficacy of algorithms? How does (or should) the type of algorithm contemplate change his analysis?*
    - *Bias In, Bias Out* by Sandra Mayson, 2019 (Read pages 2233-2249, 2262, 2277-2281)
        - *This article presents an overview of commonly contemplated metrics in the algorithmic fairness literature. Which metrics do you find most appropriate? Does it depend on context, and if so how? Do you agree with characterizing system design as policy? How does this clash with the way we generally conceive of policy?*
    - "2023 Year End Report of the Federal Judiciary" by Chief Justice John Roberts, 2023 (Read pages 1-7)
        - *This piece reflects on the way ML may fit into the judiciary. What are the applications the Chief Justice views as especially promising? What baselines should we look to when considering whether to roll out ML? Are there frameworks we can apply to make sure such systems are living up to their promise?*
    - "Inherent Limitations of AI Fairness" by Maarten Buly & Tijl De Bie, 2024
        - *This article details the inherent limitations of technical interventions on the fairness space. Are there contexts in which you think technical interventions can categorically never succeed?*

Week 6.5: Fine Tuning Demonstrations (Friday 03/01, 12:30 - 2pm via Zoom)

- **Discussion leaders:** Josh Joseph and Tom Zick
- This session will be Zoom-only and recorded for anyone who cannot attend
- Zoom Link: https://harvard.zoom.us/j/95591972511
    - Recording can be found under week 6.5 on Canvas > Syllabus

Week 7: Alignment (03/04 & 03/05)

How do we better understand two different types of alignment issues: 1) misuse by the second party (bomb making instructions, adversarial examples) vs 2) abuse of second party (biased /incorrect answers)?

- **Case:** Alignment applied within SOTA LLM (OpenAI's ChatGPT)
    - **Guest:** Tyna Eloundou (03/05)
        - *"Tyna Eloundou is a policy researcher at OpenAI who has worked on a range of topics including economic impact analysis, alignment targeting, and safety analysis."*

- **Assignment:**
    - **GROUP 4:** Firestarter Group Assignment + Individual Posts due Wednesday, 3/06 by 11:59pm. All other students expected to engage in discussions.

- **Readings:**
    - Monday
        - *Constitutional AI: Harmlessness from AI Feedback*, by Yuntao Bai et al, 2022 (Read pages 1 - 6, skim 20-22 (starting at C.), lightly skim 23 - 28)
            - *This paper describes how Anthropic (a major LLM provider) thinks about aligning their AI systems. What sources do you think are important to include in the system's 'constitution'? What concerns do you have about using an AI system to tune another AI system's helpfulness and harmfulness?*
        - "Synthetic Data: Anthropic's CAI from fine-tuning to pretraining, OpenAI's Superalignment, tips, types and open examples" by Nathan Lambert, 2023
            - *Synthetic data is an important piece of training for modern LLMs. What concerns do you have about using an AI system to generate synthetic data that is then used to train other AI systems?*
        - "How OpenAI is approaching 2024 worldwide elections" by OpenAI, 2024
            - *How OpenAI thinks about protecting the 2024 elections. What additional measures do you think are important for OpenAI to put in place?*
        - *Adversarial Examples are Not Bugs, They are Features* by Andrew Ilyas et al, 2019 (Read Section 1 and skim for anything else you find interesting!)
            - *Adversarial attacks are examples of a way of exploiting the "brittleness" of neural networks. For applications like image classification or autonomous driving, what safeguards should be put into place knowing that these networks are vulnerable to these types of attacks?*
    - Tuesday
        - *Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback* by Hannah Rose Kirk, Bertie Vidgen, Paul Rottger, and Scott Hale, 2023
            - *This article presents a taxonomy of risks and benefits from personalizing LLMs and a framework for governing them. Where do you think the lines should be drawn between the policy framework tiers and who should be able to decide that?*
        - "Democratic inputs to AI grant program: lessons learned and implementation plans" by Tyna Eloundou and Teddy Lee, 2024 (skim the report for a project you find interesting)
            - *An overview of research OpenAI is funding to decide what rules AI systems should follow. What project is most interesting to you? What do you think the process should look like?*
        - Optional (*why we might be skeptical about attempts at alignment/fundamental intuition for why alignment is difficult):*
            - "The Waluigi Effect" by Cleo Nardo, 2023
                - *A description of one of the challenges when attempting to align a LLM. Say these issues are impossible to eliminate. How should we best mitigate the presence of "Waluigis"?*

- *Practices for Governing Agentic AI Systems* by Yonadov Shavit et al, 2023
    - *A proposal of a definition for "agentic AI systems" and practices for governing them. Pick a couple open questions from the document. How would you answer those questions?*

_____ SPRING BREAK: Week of March 11_____

Week 8: Social Network Exploitation & Moderation (03/18 & 03/19)

Can algorithms be beneficial to social networks? What are potential issues?

- **Case:** Super Human Persuasion
    - **Guest:** Dr. Rumman Chowdhury (03/18)
        - *"Dr. Rumman Chowdhury's passion lies at the intersection of artificial intelligence and humanity. She is a pioneer in the field of applied algorithmic ethics, creating cutting-edge socio-technical solutions for ethical, explainable and transparent AI. Dr. Chowdhury currently runs Parity Consulting, Parity Responsible Innovation Fund, and is a Responsible AI Fellow at the Berkman Klein Center for Internet & Society at Harvard University. She is also a Research Affiliate at the Minderoo Center for Democracy and Technology at Cambridge University and a visiting researcher at the NYU Tandon School of Engineering. Previously, Dr. Chowdhury was the Director of META (ML Ethics, Transparency, and Accountability) team at Twitter, leading a team of applied researchers and engineers to identify and mitigate algorithmic harms on the platform. Prior to Twitter, she was CEO and founder of Parity, an enterprise algorithmic audit platform company. She formerly served as Global Lead for Responsible AI at Accenture Applied Intelligence. In her work as Accenture's Responsible AI lead, she led the design of the Fairness Tool, a first-in-industry algorithmic tool to identify and mitigate bias in AI systems."*

- **Assignment:**
    - **GROUP 5:** Firestarter Group Assignment + Individual Posts due Wednesday, 3/27 by 11:59pm. All other students expected to engage in discussions.
    - **Paper Assignment #2** due ~~Tuesday, 3/19 at 11:59pm~~ Monday, 3/25 at 12:00pm on Canvas

- **Readings:**
    - Monday
        - Twitter's Recommendation Algorithm, by Twitter, 2023

        - *This post shares details regarding how the company has built the social platform's algorithm (when it was still Twitter) and what features are core to its design.*
    - "Musk's own platform: Twitter's algo tracks how Elon Musk's tweets are doing, boosts them often" by Mehul Reuben Das, 2023
        - *This piece shares about how, in the Elon Musk-Twitter, a portion of the open source code revealed that Twitter tracks and boosts Elon Musk's tweets. The source code also revealed that Twitter tracked tweets about Democrats and Republicans as well.*
    - Tweet, by @Sandeep, 2023
        - *This tweet shows a snapshot of Twitter's open source code that tracks and promotes particular individuals.*

- Tuesday
    - "Sam Altman Warns That AI Is Learning Superhuman Persuasion", by Maggie Harrison, 2023
        - *This article dives into the infamous tweet by Sam Altman stating what the future of AI might look like. This tweet is substantial because it left many speculating what might be on the horizon for the company and the industry.*
    - Ifs Ands or Bots, by Jonathan Zittrain, 2023 [on Canvas]
        - *This article sheds a light on what could become of social media spaces given the rise of AI.*
    - *The Coming AI Hackers*, by Bruce Schneier, 2021
        - *Within this essay, Schneier explores the implications of AI hackers, including how AI systems will be used to hack us, as well as other systems throughout society. What did Schneier miss within his analysis? How satisfying did you find his solutions?*
    - *Algorithmic content moderation: Technical and political challenges in the automation of platform governance* by Gorwa, R., Binns, R., & Katzenbach, C. *Big Data & Society*, 2020
        - *This article shares how algorithmic moderation systems actually tend to exacerbate the moderation issues rather than relieve them for most of the major social networks. What features do you see are missing from these systems to make them more effective? Do you think they can ever be effective?*

Week 9: Open Source AI (03/25 & 03/26)

What is the case for or against open source AI?

- **Case**: LLAMA 2
    - **Guest:** Anne Neuberger (03/25)
        - *"As the Deputy National Security Advisor for Cyber and Emerging Tech, Anne Neuberger serves as an advisor to the President on matters related to cybersecurity, digital innovation, and emerging technologies. She coordinates the interagency response to cyber threats and engages with*

allies and partners on cyber cooperation. With over 25 years of experience in the government and private sector, Anne brings a unique perspective and experience to this work, which is primarily around advancing US national security interests, enhancing cyber resilience, and fostering innovation and collaboration between the private and public sectors. Prior to joining the White House, she led the establishment of the NSA's Cybersecurity Directorate, bringing together thousands of intelligence analysts, cybersecurity professionals, cryptographers, researchers, and technologists. Additionally, she previously led NSA's global intelligence operations, and served as a White House Fellow."

- **Guest:** Nathan Lambert (03/26)
  - *"Nathan Lambert is a Research Scientist at the Allen Institute for AI focusing on RLHF. Previously, he helped build an RLHF research team at HuggingFace. He received his PhD from the University of California, Berkeley working at the intersection of machine learning and robotics. He was advised by Professor Kristofer Pister in the Berkeley Autonomous Microsystems Lab and Roberto Calandra at Meta AI Research. He was lucky to intern at Facebook AI and DeepMind during his Ph.D. Nathan was was awarded the UC Berkeley EECS Demetri Angelakos Memorial Achievement Award for Altruism for his efforts to better community norms."*

- **Assignment**:
  - **GROUP 3:** Firestarter Group Assignment + Individual Posts due Wednesday, 3/27 by 11:59pm. All other students expected to engage in discussions.
  - **Paper Assignment #2** due Monday, 3/25 at 12:00pm on Canvas

- **Readings:**
  - Monday:
    - *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* By Xiangyu Qi et al, 2023
      - *This article explores the drawbacks of using certain sets of data to fine tune and align models. It brings up an idea regarding if we know locksmiths exist and could open any door any time without a key, why do we still lock our doors? Does this concept translate over to the open source debate?*
    - *Dual use of artificial-intelligence-powered drug discovery* by Fabio Urbina et al, 2022
      - *Authors wondered about the possibility of repurposing their commercial pharmaceutical company to instead build biochemical weapons. They did an experiment and realized it works. What warnings does this create for you about how to move forward with the AI industry?*
  - Tuesday:
    - "Open LLM Company Playbook", by Nathan Lambert, 2023

        - *Lambert makes a business case for open LLM weights. He outlines three requirements, actions, and benefits, along with additional notes about the future of these open models. What stood out to you, and do agree with moving more toward open weight models?*
    - "We need to push the notion that only open source LLMs can be "safe"" by @meghan_rain on Hacker News, 2023
        - *This page is a thread of people debating existential risk and open source. As you read, think through different types of ex-risk and how such conceptions should influence governance decisions.*
    - Optional:
        - "The koan of an open source LLM," by Nathan Lambert, 2024

Week 10: From the Digital to the Physical (04/01 & 04/02)

What responsibilities should attach to technologies that people rely on to understand and navigate the immediate physical world?

- **Case:** Aira
    - **Guests for 4/01:**
        - Kyle Keane @ MIT
            - *Kyle Keane is currently a Lecturer at Massachusetts Institute of Technology (MIT) in the Department of Materials Science and Engineering. He teaches computational materials science, human-centered design, and engineering technology that helps people with disabilities. He directs an undergraduate research group called the Interactive Materials Education Laboratory (IMEL) where students use technology and computation to make science more tangible and creatively engaging.*
        - Troy Otillio, CEO @ Aira
            - *Troy Otillio is a Chief Executive Officer at Aira. Mr. Otillio is a noted technology leader and entrepreneur whose lengthy experience in the industry has included managing successful startups and spearheading the first enterprise-scale adoption of AWS infrastructure at Intuit, Inc., At Aira, Troy focuses on scaling Aira's business functions,including marketing, sales, engineering and agent services. He is also actively seeking talent and collaborators in the areas of AI, NLP, IoT, Marketing, QA, and Security. Troy is an initial investor and Advisor with Aira with a tenure stretching back to the inception of the company. With over 25 years of experience, Otillio is a software product and technology entrepreneur having particular expertise in early-stage successful startups (including Ariba and Documentum), and large high-tech enterprises such as Intuit, where he served as Director of Public Cloud in the company´s Central Technology organization. Still, who earned his degree in*

computer science from California Polytechnic State University-San Luis Obispo and holds four patents in technology, also served as Executive in Residence at Plug and Play Tech Center, a Silicon Valley-based accelerator specializing in emerging tech startups.

- Geoffrey Peddle, CTO @ Aira
  - *Geoffrey Peddle has a diverse and extensive work experience. Geoffrey is currently serving as the Chief Technology Officer at Aira since May 2023. Prior to this, they worked as the Senior Director of Data Engineering at BenchSci from December 2022 to May 2023. Before that, they held the position of VP of Engineering at Seashell from November 2021 to June 2022. Geoffrey has also worked at tealbook - Instant Qualified Supplier Identification, where they served as the CTO from January 2017 to May 2021. During their tenure, they led multiple teams, including engineering, machine learning, data, and product, and played a significant role in transitioning the company from an application company to a data company. Before joining tealbook, Geoffrey was the V.P. R&D at Riva Modeling Systems from December 2013 to May 2016. Geoffrey also held positions at Personagraph as the Chief Architect, Chief Scientist, and Director of Inference Science from November 2012 to December 2013. Earlier in their career, Geoffrey worked as an Independent Consultant at Good enough innovations from January 2012 to October 2012. Geoffrey also gained research experience as a Visiting Student Researcher at Chango in September 2011 and at Google from May 2011 to August 2011.*

- **Assignment:**
  - **GROUP 6:** Firestarter Group Assignment + Individual Posts due Wednesday, 4/03 by 11:59pm. All other students expected to engage in discussions.
  - **Paper Assignment #3** due Monday, 4/22 at 12:00pm on Canvas
    - For the final short paper assignment, we invite you to draw from what you have learned and thought about over the last twelve weeks and suggest an intervention of your own design (technical, policy/legal, or your desired mix) for any of the cases we've covered. Please include a description of what your intervention would entail, and how it might come about, including barriers and how they might be overcome.
    - We invite you to think big, drawing on the actual structures you would need to mobilize to get your intervention done. Please include references to course readings and discussions as relevant, as well as whether something like this intervention has been attempted before and what your changes would be if so.

- **Readings:**
  - Monday
    - *An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility* by Kate Glazko et al, 2023.
      - *This piece is important because it discusses personal and professional uses of generative AI for accessibility. What experiences have you had with generative AI tools that may be helpful for someone with a disability? What challenges might they confront with attempting to use the generative AI tool?*
    - "AI Could Change How Blind People See the World" by Khari Johnson, 2023
      - *This piece discusses the uses and challenges of GPT-4 using to help people who are visually impaired. What, if any, responsibility do AI companies or services have to their users who may learn to rely on them to navigate their world?*
    - "Personalized ASR Models from a Large and Diverse Disordered Speech Dataset" by Katrin Tomanek, Software Engineer and Bob MacDonald, 2021
      - *Automatic speech recognition systems are being widely deployed but are challenging for people with speech impairments to use. To what degree should products with consumer speech recognition systems be robust to speech impairments? How should just a requirement be implemented?*
    - "Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities" by Ben Hutchinson et al, 2019
      - *This paper discusses disability-related biases in AI models and data. How would you propose remedying the issues highlighted in this paper? What is the primary challenge to implementing your remedies?*
    - "Disability rights advocates are worried about discrimination in AI hiring tools" by Sheridan Wall & Hilke Schellmann, 2021
      - *This article discusses the issues around the use of AI in hiring, particularly discrimination against people with disabilities. What regulation and auditing is needed to use AI tools such as the ones discussed in the article?*
    - "How Kathryn Webster Owns Her Own Story, and Her Advice for Other Blind and Low Vision Employees" by Aira Communications, 2022
      - *This blog post from Aira (an upcoming guest of ours) describes an HBS student's experience using their service which connects blind and low vision individuals with visual interpreters. What concerns would you have about using a visual interpreter like Aira provides? How would your concerns change if it were an AI system rather than a person providing the visual interpretation?*
  - Tuesday
    - Explore Sora:
      - Sora
      - Video generation models as world simulators

        - 24 Sora examples from Twitter/X that are not in OpenAI's Sora webpage : r/MediaSynthesis
            - *Sora, a video generation model recently developed by OpenAI, not only produces state-of-the-art results but is claimed to be a simulator of the physical world. Do you think Sora is a general purpose simulator of the physical world? What properties would be necessary for it to be?*
    - World Model LLM Debate:
        - Yann LeCun's Post on X, 2024
        - Jim Fan's Post on X, 2024
        - "Muddles about Models" by Gary Marcus, 2023
            - *These readings highlight the significant disagreement on if and to what degree it can be said that LLMs or video generation models contain a model of the world. Which of these arguments do you find most compelling? Why*
    - "OpenAI Sprinting to Keep Up With Startups on AI-Generated Video" by Rachel Metz, 2024
        - *This article discusses the companies working on AI video generation products and the disinformation challenges they pose. Do you agree with Giada Pistilli's that the downsides outweigh the positives?*
    - "Are Video Generation Models World Simulators?" by Raphael Milliere, 2024
        - *This article presents a nuanced analysis of the claim that Sora is a "world simulator." Feel free to simply skim the article in those places where it delves into the technical. If we could rigorously show that Sora or a LLM contained a world simulator what would be the implications for trust and safety?*
    - *Dissociating language and thought in large language models* by Kyle Mahowald et al, 2023
        - *This paper introduces the distinction between formal linguistic competence and functional linguistic competence in order to understand language-thought conflation fallacies that arise from the development of LLMs. We offer this article here so that you can get a flavor of how scientists – in this case, cognitive scientists – are trying to make sense of how LLMs work, what they do, and how "smart," in fact, they are. What tasks do you regularly encounter where high formal linguistic competence but low functional linguistic competence is sufficient for an AI system to be useful to you?*
    - "LLMs differ from human cognition because they are not embodied." by Anthony Chemero, 2023
        - *This article articulates the differences between LLMs and human cognition. Do you believe that embodiment is necessary to "give a damn"?*

Week 11: No Class Monday and Tuesday (04/08 & 04/09)

- No Assignment - work on Paper Assignment #3 due Monday 4/22 at 12:00pm on Canvas

    - For the final short paper assignment, we invite you to draw from what you have learned and thought about over the last twelve weeks and suggest an intervention of your own design (technical, policy/legal, or your desired mix) for any of the cases we've covered. Please include a description of what your intervention would entail, and how it might come about, including barriers and how they might be overcome.
    - We invite you to think big, drawing on the actual structures you would need to mobilize to get your intervention done. Please include references to course readings and discussions as relevant, as well as whether something like this intervention has been attempted before and what your changes would be if so.

Week 12: Existential Risk & AGI (04/15 & 04/16)

What is the best case for concern about Existential Risk of Artificial Intelligence? What ethical considerations arise when discussing existential risks of AI, and how can we balance technological progress with safeguarding humanity's future?

- **Case:** Slaughterbots (pre-LLMs vision of x-risk)

- **Assignment:**
    - **GROUP 2:** Firestarter Group Assignment + Individual Posts due Wednesday, 4/17 by 11:59pm. All other students expected to engage in discussions.
    - **Paper Assignment #3** due Monday, 4/22 at 12:00pm on Canvas
        - For the final short paper assignment, we invite you to draw from what you have learned and thought about over the last twelve weeks and suggest an intervention of your own design (technical, policy/legal, or your desired mix) for any of the cases we've covered. Please include a description of what your intervention would entail, and how it might come about, including barriers and how they might be overcome.
        - We invite you to think big, drawing on the actual structures you would need to mobilize to get your intervention done. Please include references to course readings and discussions as relevant, as well as whether something like this intervention has been attempted before and what your changes would be if so.

- **Readings:**
    - Monday
        - *The Vulnerable World Hypothesis*, by Nick Bostrom, 2019
            - *Bostrom asks us to consider what it would look like to discover a technological "black ball" and argues that given the "semi-anarchic" condition of global government we have a "vulnerable world." Do you agree with what Bostrom thinks are the salient characteristics of governance? Which vulnerability do you find most compelling? Do hiss stabilization suggestions resonate with you?*

        - *Minds, Brains, and Programs*, by John R. Searle, 1980 [Skip the replies, eg read 1-5 & 10-14]
            - *Searle advances a theory that only machines with internal causal powers equivalent to those of brains can think. Since AI is about programs and not machines, Searle argues it cannot think. Do you find this reasoning convincing? If not, is the nature of Searle's equivocation logical? technical?*
        - *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, by Sebastian Bubeck et al., 2023 [Read Introduction & Conclusion, skim rest for interest]
            - *This paper aims to track the axes along which GPT4 amounts to a technical leap. Do you agree with the author's characterization general intelligence? How do these sparks of AGI compare to the notions of super-intelligence put forward by Searle and Bostrom?*
        - *Emergent analogical reasoning in large language models* by Taylor Webb et al, 2023. [Read intro and results only]
            - *This piece examines pattern induction in GPT-3/4 by looking at various analogy tests. Their results indicate an emergent ability to reason by analogy that sometimes outperforms humans. Are they asking the right questions about intelligence?*
    - Tuesday
        - *Overly Intelligent AI; Human Compatible*, by Stuart Russell, 2019 [Chpt 5]
            - *This chapter traces the intellectual history of existential risk in AI from its roots in the 1800s. How has this intellectual trajectory shaped the nature of the modern existential threat discussion?*
        - "How Silicon Valley doomers are shaping Rishi Sunak's AI plans" by Laurie Clarke, 2023
            - *This article examines the corporate capture theory of existential risk. To what extent should the corporate incentives at play affect our analysis of what may be real policy debates?*
        - Tweet thread by Tyler Austin Harper, https://twitter.com/tyler_a_harper/status/1726286339258429663?s=46&t=mXFpnNQL1KEyLxgRJa3GHA
            - *This thread examines the intellectual history of doom-saying and its sometimes ungenuine motives. How does this compare to the intellectual history of Ex-Risk?*