

Governing Artificial Intelligence: Law, Policy, and Institutions

Syllabus last revised: Oct 13th

Updates: Added two excused absences and late days, swapped security and labor lectures.

Law 4052 // Computer Science 283 // Communication 252A // Political Science 245B/445B

Monday/Wednesday 2:15 - 3:45pm; Optional Thursday/Friday Section 2:30-3:20pm.

Crown Building, Stanford Law School, Room 290

cs283.stanford.edu | [Course Schedule](#) | [Course Playlist](#)

Nate Persily (School of Law, Political Science, FSI, Communication)

npersily@stanford.edu

Office – SLS Neukom Building, Room N230

Office Hours: Tues 2-4PM

Rob Reich (Political Science)

reich@stanford.edu

Office – Encina Hall Central, Room 441

Office Hours: M 430-6pm, book appointment with Dominic Zappia (zappia@stanford.edu)

Anka Reuel (Computer Science)

anka.reuel@stanford.edu

Office Hours: Friday, 11am-noon, Durand Building (Room tbd)

Sanmi Koyejo (Computer Science)

sanmi@cs.stanford.edu

Office Hours: TBD, upon request

[NOTE: Syllabus details, including reading lists, are subject to change and may be updated up until two weeks before the start of a lecture]

Course Description

Artificial intelligence (AI) is rapidly reshaping industries, institutions, and societies – and its governance is one of the most urgent and complex policy challenges of our time. This course explores how societies are responding to AI through technical interventions, law, regulation, ethics, professional standards, and institutional and organizational design. Taught by a cross-disciplinary faculty team, students will engage with historical analogues, present-day legal and policy debates, technical problems, and emerging governance models across democratic, authoritarian, and multilateral contexts. The course combines philosophical, social-scientific, legal, and technical readings with real-world case studies, inviting students to explore the limits and possibilities of governing AI.

Sections / Distinguished Guests / Supplementary Workshops

- In addition to the lectures on Mondays and Wednesdays, we will also hold optional sections on Thursday (2:30-3:20) and Friday (2:30-3:20) in Room 14 in the basement of the Crown Building of the Law School.
- We will also be hosting several guest speakers outside of class throughout the term and we encourage you to attend those sessions.
- On Thursdays from 12:40 to 2:00 PM in 270 Crown (over the breezeway) the Law School's AI Initiative will have its weekly speaker series, featuring leading figures in the AI ecosystem. Lunch is provided so please RSVP every time you would like to attend and everyone should register to receive emails from the Initiative. You should sign up for emails and to RSVP for events at this link: <https://law.stanford.edu/ai-initiative/events-ai/>

Learning Objectives

By the end of the course, students will be able to:

1. Explain the technological foundations of modern AI systems, including how they are trained, evaluated, and deployed, and describe their broader societal implications.
2. Engage with technical governance problems, such as meaningful evaluation design, system access considerations, and privacy-preserving analysis of use and misuse of AI systems.
3. Critically evaluate the technical, organizational, and normative dimensions of AI oversight.
4. Analyze the historical patterns of governance responses to frontier technologies.
5. Identify and assess key legal, technical, and institutional approaches to AI governance across national and global contexts.
6. Anticipate and reason through future challenges in AI governance, including the role of closed vs open models, (multi-)agents, low-resource models, disinformation, and global coordination.
7. Synthesize interdisciplinary perspectives to formulate and evaluate actionable governance proposals.

Instructors

[Rob Reich](#) is the McGregor-Girand Professor of Social Ethics of Science and Technology at Stanford University and Associate Director of the Stanford Institute for Human-Centered AI. In 2024-25, Rob was on public service leave as Senior Advisor to the US AI Safety Institute. His scholarship in political theory engages with the work of social scientists and engineers. His newest work is on governance of frontier science and technology. His most recent books are *System Error: Where Big Tech Went Wrong and How We Can Reboot* (with Mehran Sahami and Jeremy M. Weinstein) and *Digital Technology and Democratic Theory* (edited with Lucy Bernholz and Hélène Landemore).

Nate Persily is the James B. McClatchy Professor of Law at Stanford Law School, with appointments in the departments of Political Science, Communication, and the Freeman Spogli Institute of International Studies. He is the Founding Co-Director of the Stanford Cyber Policy Center and the new Stanford Law School AI Initiative. He also co-chaired the AI Task Force of the American Political Science Association. His most recent book is a co-edited volume, *The Digitalist Papers: Artificial Intelligence and Democracy in America* (2024).

Anka Reuel is a PhD candidate in Computer Science at Stanford where she conducts technical AI governance research at the Stanford Trustworthy AI Research Lab and the Stanford Intelligent Systems Laboratory and an incoming Assistant Professor at Harvard University with appointments in Public Policy & Computer Science. She served as one of the vice chairs for the EU's first General Purpose AI Code of Practice which implements the EU AI Act and specifies concrete obligations for foundation model providers to show compliance with the AI Act. She also was a 2024-25 Technology and Geopolitics Fellow at the Belfer Center at Harvard Kennedy School. Anka is also the lead writer for the AI Chapter of the 2025 Stanford Emerging Technology Review and the Lead Researcher for the Responsible AI Chapter of Stanford's AI Index.

Sanmi Koyejo is an Associate Professor in the Department of Computer Science at Stanford University, where he leads the Stanford Trustworthy AI Research Lab. His research interests are in developing the principles and practice of trustworthy artificial intelligence and how to apply these findings to inform better policymaking. Sanmi's work has won multiple awards, including a NeurIPS 2023 best paper award for his team's work on whether emergent abilities are a mirage. This work has subsequently strongly influenced the political discourse on emergent abilities.

Course Staff

Bruria Friedman-Feldman – Law Teaching Assistant
bruriaff@stanford.edu

Tia Vasudeva – Teaching Assistant
tvasudev@stanford.edu

Yifei Cheng – Teaching Assistant
yifechen@stanford.edu

Dominic Zappia – Head Course Assistant
Office Hours: Weds 3:45-4:45PM (after class; then law school cafe)
zappia@stanford.edu

Roberta Fischli – Course Assistant
fischli@stanford.edu

Yasi Khan — Course Assistant
yrkhan@stanford.edu

Grading and Assignments

We have assigned a carefully curated set of readings for each class session. We expect you to complete this reading in advance of each class session, and to come to class prepared to discuss the materials.

Because of different rules (and grading systems) between the law school and the rest of the university, law students will have a different set of requirements from other students in the class. The law students, who will be graded by Professor Persily, may miss no more than two lectures if you wish to be eligible for Honors for the class. However, you are not expected (but are nevertheless encouraged) to attend the last class, on December 2nd, as it conflicts with the Law School's reading period. Your grade (on the H-P system and upper level class curve) will be determined by two things: a final research paper of no fewer than 25 pages on any topic within the subject matter of the class, and on class/section participation. An outline of the paper is due at 5 PM Pacific Time on October 19. The paper is due January 3 at 5 PM Pacific Time. If you are taking this class for "R" credit, the paper shall be 30 pages in length and, in addition to the above requirements, you also must submit a rough draft by November 16 at 5 PM Pacific. You may use AI tools in the research and production of those papers, but (1) you must disclose in a separate document how you used AI, and (2) any hallucinations in the paper shall result in a failing grade for the course. In addition, you must write your outline and paper in Google Docs and share the document with Professor Persily, so he can track your use of AI and the progress on your paper.

The following are the requirements for the non-law students in the course:

Assignments and Grading Breakdown

The course includes two assignments. The first assignment is a substantial research project. We've broken down the project into milestones to provide you with support and feedback along the way. We encourage you to reach out early with any questions. You will receive more information about each of the assignments well in advance of their due dates. All students begin with two late days for the entire quarter. (Not per assignment.) These two late days may be used on any assignment throughout the quarter. **All assignments are due on Canvas at 5PM PST unless otherwise stated. We will provide formatting templates for all assignments.**

- AI Evaluation Research Project
 - You will conduct a research project in any of the three proposed tracks.
 - You can work in groups of up to 2 people.
 - Milestone 1: Question, relevance, and novelty write-up – due **October 16th** at 5PM

- Students state the question they want to answer, why anyone would or should care about the answer, and conduct a thorough literature review establishing that this question is novel and has not already been sufficiently addressed. (3 to 4 pages)
- Milestone 2: Project Plan – due **October 30th** at 5PM
 - Students finalize the question, core contribution they want to make, methods, how they know if they succeeded, what resources and access they need, most likely ways it can fail and fallback strategies. (2-3 pages)
- Milestone 3: Peer Review – due **November 6th** at 5PM
 - 500-word peer review on another project plan.
- Final Paper – due **December 4th** at 5PM
 - Final research paper plus decision memo/impact plan (jargon-free, addressed to a specific office, committee, agency, or someone in industry, not "policymakers" or "labs"). (10-12 pages)

Check Canvas for more information about the final project.

- Public Comment Assignment – due on **October 23rd at 5PM**
 - Students are asked to respond to the [Request for Comments \(RfC\) on the Colorado Automated Decision-Making Technology and Chatbot Safety Acts](#).
 - You don't have to actually submit your comment (but we strongly encourage it!)
 - 1,000 (min) to 1,500 (max) word comment:
 - Depending on your opinion of the Act, you can either write about the parts that you agree with are good interventions or provide feedback on what should be changed.
 - Your commentary can, for example, focus on a proposed wording, a threshold, a definition, or a procedural step (or a mix of these)
 - Provide a rationale for why you endorse a part of the bill or why you propose a change
 - Ground your claims in evidence you can point to: data, a documented case, a technical constraint, a citation. Don't write things like "experts agree" without backing it up with sources.

Grades for non-law students will be calculated as follows:

- AI Evaluation Research Project – 60%
 - 4-page question, relevance, and novelty write-up – 25%
 - Project Plan – 20%
 - Peer Review – 10%

- Final Paper – 45%
- Public Comment Assignment – 30%
- Attendance/Participation – 10%

Tracks for AI Evaluation Research Project

Track 1: Who decides what gets measured?

Today, the organizations best positioned to evaluate frontier AI systems are the companies building them, and the evaluations that exist largely reflect what frontier labs and (academic) benchmark creators have chosen to prioritize. There is no guarantee that this portfolio matches what society most needs to know. Some capabilities and risks are measured obsessively while others, arguably more consequential, have no measurement at all, and whatever gets measured becomes a target that shapes what companies build. This track asks how evaluation agendas should be set and how the ecosystem around evaluation could be redesigned. Projects might examine who currently funds, produces, and consumes evaluations and where the gaps and conflicts of interest lie and how these are reflected in the coverage and technical design of the evaluation; what mechanisms could surface public or governmental priorities for what gets evaluated, and whether analogies from other industries (financial auditing, crash testing, clinical trial endpoints, environmental monitoring) transfer; what incentives, institutions, or market structures would sustain independent, high-quality evaluation; or how agenda-setting power is distributed between developers, evaluators, regulators, and affected communities, and what would change if it were distributed differently.

Track 2: Building evaluations that deserve trust

An evaluation is a measurement instrument, and most AI evaluations would not survive the scrutiny routinely applied to instruments in psychometrics, medicine, or engineering. Scores are reported without uncertainty, benchmarks are used far outside the conditions they were designed for, and it is often unclear what construct a benchmark actually measures. This track asks how to make evaluations more valid, reliable, and transparent, and what it would take for an eval result to constitute real evidence for deployment decisions, in court, ... Projects might diagnose the reliability or validity of an existing benchmark and what this means for using them in a specific decision-making context in the real world; develop or apply methods for quantifying and reporting uncertainty in eval results and outlining how these uncertainty bounds should be taken into account by a decision-maker; design reporting standards or documentation infrastructure that would let a third party assess whether an eval supports a given claim; or investigate a specific measurement failure mode (contamination, saturation, judge unreliability, construct drift) and what mitigations actually work.

Track 3: Evaluations at the point of decision/deployment

Evaluations only matter when someone acts on them such as a regulator clearing a model for deployment, a hospital procuring a diagnostic tool, a court weighing whether a system discriminated, a company deciding its model is safe enough to ship. This track asks what happens at that interface. How much evidence, of what quality, should be required for decisions of different stakes? How does that translate to design decisions for evaluations? Projects might analyze how eval evidence is actually used (or misused) in a real decision context such as the EU AI Act conformity regime, procurement, litigation, or lab deployment decisions; propose evidence standards or decision thresholds for a specific high-stakes context and defend them; examine what happens when eval evidence is ambiguous or contested and who bears the burden of proof; or study how eval results get communicated to decision-makers and where meaning is lost in translation and the potential risks this carries.

Course Materials

All course readings are provided in online format, accessible by any student with a SUNet ID. There are no textbooks required. Stanford University and its faculty are committed to ensuring that all courses are financially accessible to all students. If you are an undergraduate who needs assistance with the cost of supplies, course materials and/or fees, you are welcome to approach us directly.

Prerequisites

None

Lecture Attendance

Except in cases of an OAE accommodation or another valid reason that is brought to the course managers or a student's course assistant's attention *before* lecture, attendance at lectures is mandatory. If a student has a prolonged illness or a personal situation that might lead to more than one lecture absence, the student should contact a member of the course staff before missing a lecture. Attendance will be taken at the beginning of each class. Any student coming in late shall be marked as absent. *Students will be given two excused absences from lecture without penalty. These absences are designed to account for illness or other extenuating circumstances.*

In order to be prepared for in-class discussions, it is essential that you come to each lecture having read and understood the materials assigned and having given some thought as to how the readings relate to the course in general. This will allow you to benefit from the class presentations and discussions and in turn prepare yourself to discuss the issues in depth in section. You should come to class with considered views about:

- *what the main claims offered in the texts or case studies are;*
- *the arguments offered in favor of these claims;*

- *whether these are good or plausible arguments;*
- *whether the claim is, all things considered, strong or plausible;*
- *what alternatives to the claims and arguments exist; and*
- *whether some alternative is superior to the claim under discussion.*

The Honor Code

Violating the Honor Code is a serious offense, even when the violation is unintentional. Please read and review the [Honor Code here](#). You are responsible for understanding the University rules regarding academic integrity; you should familiarize yourself with the code if you have not already done so. In brief, conduct prohibited by the Honor Code includes all forms of academic dishonesty, among them copying from another student's work, unpermitted collaboration and representing as one's own work the work of another. If you have any questions about these matters, see your course assistant during office hours.

Statement on the Use of Automated Writing/Coding Tools

The development of generative AI (a topic in the class) provides opportunities to support learning but also opportunities to cheat. The use of automated writing or coding tools (e.g., ChatGPT, Gemini, or Claude) for submitting assignments is permitted in this class. For any assignment, however, students must disclose in a separate document how they used AI tools for the assignment, including a complete catalogue of prompts. In addition, any hallucinations in submitted work will result in a failing grade for the course.

Device Free Classroom

In accordance with [Stanford PACE](#), we will be adopting a device free classroom policy. Please review our policy below. If you have any questions or accommodations, please contact your teaching assistant.

Technology: To enhance your learning and the focus of those around you, laptops, tablets, smartphones, and wearable devices like smart glasses or smart watches are not permitted in class except as an accommodation for a documented disability. Please keep all devices silenced and stowed in your bag for the entire duration of class.

FERPA

Student Record Privacy Policy

<http://studentaffairs.stanford.edu/registrar/students/ferpa>

Access and Accommodations

Stanford is committed to providing equal educational opportunities for disabled students.

If you experience disability, please register with the Office of Accessible Education (OAE). Professional staff will evaluate your needs, support appropriate and reasonable accommodations, and prepare an Academic Accommodation Letter for faculty. To get started, or to re-initiate services, please visit <https://oae.stanford.edu>.

If you already have an Academic Accommodation Letter, we invite you to share your letter with us. Academic Accommodation Letters should be shared at the earliest possible opportunity so we may partner with you and OAE to identify any barriers to access and inclusion that might be encountered in your experience of this course.

Additional Resources for Learning

Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Student Disability Resource Center (SDRC) located within the Office of Accessible Education (OAE). SDRC staff will evaluate the request with required documentation, recommend reasonable accommodations and prepare an Accommodation Letter for faculty dated in the current quarter in which the request is being made. Students should contact the SDRC as soon as possible since timely notice is needed to coordinate accommodations.

Student Disability Resource Center Office of Accessible Education
<http://studentaffairs.stanford.edu/oae>

The Hume Writing Center works with Stanford students taking WIM classes and any course that includes writing assignments. In free one-to-one sessions, trained writing consultants help students brainstorm and get started on assignments; learn strategies for revising, editing, and proofreading; and improve organization, flow, and argumentation. We also have digital media consultants who work with students to develop strategies to improve visual and multimodal communication in media such as research posters and PowerPoint and oral communication tutors to help students prepare or refine a presentation. Students can make an appointment with a lecturer or advanced graduate student consultant or drop in to meet with an undergraduate peer tutor. For further information, to see hours and locations, or to schedule an appointment, visit the Hume website at: <http://hume.stanford.edu>, <http://hwc.stanford.edu/>

The Technical Communication Program (TCP) is a writing and public speaking resource for Stanford students of all levels. TCP is a resource specifically tailored for WIM courses offered in the School of Engineering. Mary McDevitt, Ph.D. (mary.mcdevitt@stanford.edu), Director of TCP, is available to all students for consultation. TCP instructors will also be working with WIM students to provide support on writing assignments. For more information, please visit: <https://engineering.stanford.edu/tcp>.

Schedule

Week 1: Foundations of AI and the Governance of Frontier Technologies

- Weds, Sep 23 Lecture 1: Introduction: Methodological Approaches & Competing Visions of the AI Transition
- Week 2: Technical Foundations
 - Mon, Sep 28 Lecture 2: Technical Foundations & Lessons from the History of Governing Frontier Technology
 - Weds, Sep 30 Lecture 3: AI Agents
- Week 3: Technical Foundations & Individual Risks
 - Mon, Oct 5 Lecture 4: AI Evaluations
 - Weds, Oct 7 Lecture 5: Bias, Discrimination, and Fairness
- Week 4: Individual & Societal AI Risk
 - Mon, Oct 12 Lecture 6: AI, Copyright, Creativity & Privacy
 - Weds, Oct 14 Lecture 7: AI Harms: Defamation & Mental Health
- Week 5: Societal Risk from AI cont.
 - Mon, Oct 19 Lecture 8: AI & Democracy
 - Weds, Oct 21 Lecture 9: AI Infrastructure & the Environment
- Week 6: Existential Risk & Policy
 - Mon, Oct 26 Lecture 10: Existential Risks & Benefits, AGI, Autonomous R&D, CBRN
 - Weds, Oct 28 Lecture 11: EU Law and Policy
- Week 7: US & Global AI Governance Policy
 - Mon, Nov 2 Lecture 12: Federal & State-Level US Law and Policy
 - Weds, Nov 4 Lecture 13: China and Middle Powers Approaches to AI Governance
- Week 8: The AI Economy
 - Mon, November 9 Lecture 14: The AI Economy: The Future of Labor
 - Weds, Nov 11 Lecture 15: The AI Economy: Competition, Open Models, & Antitrust
- Week 9: Security & Human Futures
 - Mon, Nov 16 Lecture 16: Geopolitics and Global AI Governance
 - Weds, Nov 18 Lecture 17: What Does It Mean to Be Human in an AI Age?
- Week 10: Outlook & AI Futures
 - Mon, Nov 30 Lecture 18: TBD
 - Weds, Dec 2 Lecture 19: The Future of AI Governance

Syllabus

Week 1: Foundations of AI and the Governance of Frontier Technologies

- **Weds, Sep 23 Lecture 1: Introduction: Methodological Approaches & Competing Visions of the AI Transition**

Summary: Introductions from the instructors, with reflections on their different disciplinary and methodological approaches to AI governance. This session will explore the diverse narratives shaping the public and policy discourse on AI, from techno-optimist manifestos and visions of AGI and ASI to perspectives on AI as a “normal” technology and arguments for managing extreme risks.

Required Readings:

- Altman, S. (2024). [The Intelligence Age](#). (3 pages)
- Amodei, D. (2026), [We Must Pace the Frontier](#). (20 pages)
- Andreessen, M. (2023), [The Techno-Optimist Manifesto](#), Andreessen Horowitz. (18 pages).
- Kokotajlo, D., Larsen, T., Lifland, E., Dean, R., Greenblatt, R., & Halstead, B. (2026), [AI 2040: Plan A](#). AI Futures Project. (90 pages)
- Toner, H. (2025), [Unresolved debates about the future of AI](#), Rising Tide. (20 pages)

Supplementary:

- Amodei, D. (2024), [Machines of Loving Grace](#). (46 pages)
- Bengio, Y. (2026). [International AI Safety Report](#) (Executive Summary) (3 pages)
- Bengio, Y., et al. (2024), [Managing Extreme AI Risks Amid Rapid Progress](#), Science. (6 pages)
- Narayanan, A., & Kapoor, S. (2025), [AI as Normal Technology](#) (Knight Foundation). (41 pages)

Week 2: Technical Foundations

- **Mon, Sep 28 Lecture 2: Technical Foundations & Lessons from the History of Governing Frontier Technology**

Summary: This lecture provides a technical foundation for understanding how modern AI systems work. We’ll cover the core concepts that underlie today’s large-scale models, including stochastic gradient descent, neural networks, and key optimization techniques. Students will learn about model architectures used in foundation models, as well as common strategies for improving performance such as fine-tuning, quantization, and distillation. We’ll examine how scale affects model capabilities, and introduce reasoning phenomena such as chain-of-thought reasoning. This technical grounding is essential for analyzing later governance challenges related to safety, evaluation, and access.

Readings:

- Karpathy, A. (2023). [Intro to large language models](#). (YouTube video, 1 hr)

- Lages, J. (2023). [Direct preference optimization \(DPO\) — A simplified explanation](#). Medium. (6 pages)
- Lee, H., et al. (2024). [RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback](#). Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235, 26874–26901. (9 pages)
- Toner, H. (2023). [What are generative AI, large language models, and foundation models?](#) Center for Security and Emerging Technology. (3 pages)

Supplementary:

- Wolfram, S. (2023). [What is ChatGPT doing ... and why does it work?](#) Stephen Wolfram Writings. (75 pages)

● **Weds, Sep 30 Lecture 3: AI Agents**

Summary: We investigate agentic systems with memory, tool use, and goal-setting capabilities. Topics include reinforcement learning from AI feedback (RLAIF), planning in multi-agent environments, and challenges from open-ended exploration. We will discuss the governance implications of decentralized, interacting agents and some of the recent incidents that happened.

Readings:

- Gardhouse, K., & Oueslati, A. (2026). [The EU AI Act is not ready for agents](#). Tech Policy Press. (7 pages)
- Freedman, D. (2026). [Anatomy of an autonomous attack: 5 alarming A.I. capabilities](#). The New York Times. (9 pages)

Supplementary Readings:

- METR (2026). [Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident](#). (91 pages)
- Durante, Z., et al. (2024). [Agent AI: Surveying the horizons of multimodal interaction](#). (54 pages)
- Hammond, L., et al. (2025). [Multi-agent risks from advanced AI](#). Cooperative AI Foundation, Technical Report #1. (52 pages)
- Kapoor, S., et al. (2024). [AI agents that matter](#). (12 pages)
- Kolt, N. (2026). [Governing AI agents](#). Notre Dame Law Review, 101. (48 pages)
- Wang, D., et al. (2024). [Retrieve what you need: A mutual learning framework for open-domain question answering](#). Transactions of the Association for Computational Linguistics, 12, 247–263. (17 pages)

Week 3: Technical Foundations & Individual Risks

- **Mon, Oct 5 Lecture 4: AI Evaluations**

Summary: This lecture focuses on the role that evaluations play in AI governance. We explore how model evaluations are used to inform access policies, reporting requirements, and system release decisions. We'll also discuss technical challenges such as the absence of shared standards for benchmark development, concerns about (construct) validity, and data contamination in test sets. We'll further examine recent proposals for evaluation frameworks, safe harbor mechanisms for auditing, and how red-teaming practices can supplement formal evaluation protocols.

Supplementary Readings:

- Boubdir, M., et al. (2024). [Elo uncovered: Robustness and best practices in language model evaluation](#). Conference on Neural Information Processing Systems (NeurIPS). (13 pages)
- Deng, C., et al. (2024). [Investigating data contamination in modern benchmarks for large language models](#). North American Chapter of the Association for Computational Linguistics (NAACL). (10 pages)
- Hadfield, G., & Clark, J. (2023). [Regulatory markets: The future of AI governance](#). arXiv preprint. (24 pages)
- Longpre, S., et al. (2024). [A safe harbor for AI evaluation and red teaming](#). International Conference on Machine Learning (ICML), PMLR 235. (10 pages)
- Reuel, A., et al. (2024). [BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices](#). Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track. (11 pages)
- Salaudeen, O., et al. (2025). [Measurement to meaning: A validity-centered framework for AI evaluation](#). arXiv preprint. (52 pages)

- **Weds, Oct 7 Lecture 5: Bias, Discrimination, and Fairness**

Summary: We dive into technical definitions of fairness and examine algorithmic bias through a computational lens. Students will learn about fairness metrics (e.g., demographic parity, equalized odds), how bias can emerge from training data, and how interventions such as reweighting or post-processing affect outcomes. The session also explores limitations of current fairness audits and where technical and legal frameworks diverge. We will further discuss how fairness concerns intersect with civil rights law, public trust, and social legitimacy.

Readings:

- Barocas, S., Hardt, M., & Narayanan, A. (2023). [Fairness and machine learning: Limitations and opportunities](#) (chapters 1, 4, and 8). MIT Press. (71 pages)
- Gillum, R., Leslie, G., & Wong, C. (2026). [AI, race, and politics](#). In N. Persily & J. A. Tucker (Eds.), *Artificial intelligence, politics, and political science* (pp. 152–196).

APSA Presidential Task Force on Artificial Intelligence. Cambridge University Press. (45 pages)

- Lawyers' Committee for Civil Rights Under Law. (2025). [AI Civil Rights Act](#) (homepage plus [1-page summary](#)).
- Wang, A., et al. (2025, February). [Fairness through difference awareness: Measuring desired group discrimination in LLMs](#). (9 pages)

Supplementary Readings:

- Ho, D. E., & Xiang, A. (2020). [Affirmative algorithms: The legal grounds for fairness as awareness](#). University of Chicago Law Review Online. (21 pages)
- Huq, A. Z. (2019). [Racial equity in algorithmic criminal justice](#). Duke Law Journal, 68(6), 1043–1134. (92 pages)

Week 4: Individual & Societal AI Risk

- **Mon, Oct 12 Lecture 6: AI, Copyright, Creativity & Privacy**

Summary: This lecture examines how AI model training intersects with copyright law, privacy, and human creativity. The lecture focuses on disputes like *OpenAI vs. The New York Times* to explore broader themes. We'll explore distinctions between output similarity and the use of copyrighted training data, and assess potential legal consequences for model developers. Technical issues such as dataset curation, memorization, unlearning, and output replication will be discussed alongside emerging attribution mechanisms like output tracing. We will also discuss the challenges of maintaining the right to privacy with large language models.

Readings:

- Congressional Research Service. (2025, June). [Generative artificial intelligence and copyright law](#) (Legal Sidebar No. LSB10922). (7 pages)
- Cooper, A. F., et al. (2024). [Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice](#). (24 pages)
- Cooper, A. F., & Grimmelmann, J. (2025). [The files are in the computer: On copyright, memorization, and generative AI](#). Chicago-Kent Law Review, 100(1), 141–217. (76 pages)
- Ippolito, D., et al. (2023). [Preventing verbatim memorization in language models gives a false sense of privacy](#). International Natural Language Generation Conference (INLG). (9 pages)
- Lemley, M. (2024). [How generative AI turns copyright law upside down](#). Columbia Science and Technology Law Review, 25(2). (23 pages)
- Liu, K. Z., et al. (2025). [Language models may verbatim complete text they were not explicitly trained on](#). International Conference on Machine Learning (ICML). (9 pages)

Privacy Readings:

- Burgess, M. (2023). [ChatGPT has a big privacy problem](#). WIRED. (8 pages)
- European Data Protection Board. (2024, December 17). [Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models](#). (35 pages)
- noyb. (2024, April 29). [ChatGPT provides false information about people, and OpenAI can't correct it](#). (3 pages)
- noyb. (2024, June 6). [noyb urges 11 DPAs to immediately stop Meta's abuse of personal data for AI](#). (5 pages)
- Solove, D. J. (2025). [Artificial intelligence and privacy](#). Florida Law Review, 77(1), 1. (73 pages)

- **Weds, Oct 14 Lecture 7: AI Harms: Defamation & Mental Health**

Summary: We will survey the state of defamation law with respect to AI models. Students will explore attempts to understand the emerging power of AI models within a broader legal framework. This session will also cover AI, well-being, and mental health, investigating the potential harms from human chatbot interactions.

Readings:

- Bensinger, K. (2025, November 12). [Who pays when AI is wrong](#). The New York Times.
- Chatterjee, R. (2025, September 19). [Their teenage sons died by suicide. Now they are sounding the alarm about AI chatbots](#). NPR.
- Garcia v. Character Technologies, Inc., [Complaint](#) (M.D. Fla. Oct. 23, 2024). (126 pages; just read the summary of claims)
- Raine v. OpenAI, Inc., [First Amended Complaint](#) (Cal. Super. Ct. 2025). (53 pages — focus on the introductory factual allegations)
- Volokh, E. (2023). [Large libel models? Liability for AI output](#). Journal of Free Speech Law, 3(2). (70 pages)

- **Fri, Oct 16: Assignment due: Milestone 1: Question, relevance, and novelty write-up**

Week 5: Societal Risk from AI cont.

- **Mon, Oct 19 Lecture 8: AI & Democracy**

Summary: This session examines the new and emerging challenges and possibilities of AI and democracy. We explore the wide-ranging impacts of AI on elections, political movements, and the online information ecosystem.

Readings:

- Fung, A., et al. (2026). [Artificial intelligence & democracy: Campaigns, elections, movements, and deliberation](#). In N. Persily & J. A. Tucker (Eds.), *Artificial*

intelligence, politics, and political science (Chapter 1). APSA Presidential Task Force on Artificial Intelligence. Cambridge University Press.

- Nyhan, B., et al. (2026). [Easy to produce, hard to persuade: The asymmetric effects of AI on the online information ecosystem](#). In N. Persily & J. A. Tucker (Eds.), *Artificial intelligence, politics, and political science (Chapter 2)*. APSA Presidential Task Force on Artificial Intelligence. Cambridge University Press.
- Persily, N. (2024). [Misunderstanding AI's democracy problem](#). The Digitalist Papers. (17 pages)

- **Weds, Oct 21 Lecture 9: AI Infrastructure & the Environment**

Summary: In this session, we investigate the physical infrastructure underpinning frontier AI, including the massive energy and data center demands, and the resultant environmental impact. In addition, we're covering what's known as compute governance – i.e., the institutional arrangements, regulatory instruments, and technical mechanisms that structure access to, and oversight of, computational resources in the context of AI.

Readings:

- Epoch AI. (2025). [How much energy does ChatGPT use?](#) Gradient Updates. (16 pages)
- Gorey, J. (2025). [Data drain: The land and water impacts of the AI boom](#). Land Lines, Lincoln Institute of Land Policy. (18 pages)
- Institute for AI Policy and Strategy (IAPS). (2025). [How AI chips are made](#). (6 pages)
- Kneese, T., & Young, M. (2024). [Carbon emissions in the tailpipe of generative AI](#). Harvard Data Science Review. (5 pages)
- Luccioni, S., et al. (2024). [The environmental impacts of AI — Policy primer](#). Hugging Face. (11 pages)
- Masley, A. (2025). [The AI water issue is fake](#). (50 pages)
- Patel, D., Nishball, D., & Eliahou Ontiveros, J. (2024). [AI datacenter energy dilemma — Race for AI datacenter space](#). SemiAnalysis. (37 pages)

Supplementary:

- International Energy Agency. (2026). [Key questions on energy and AI](#). IEA, Paris. (138 pages)
- Schmidt, E. (2025). Testimony before the House Committee on Energy and Commerce, [Converting energy into intelligence: The future of AI technology, human discovery, and American global competitiveness](#). (11 pages)

- **Fri, Oct 23: Assignment due: Public Comment Assignment**

Week 6: Existential Risk & Policy

- **Mon, Oct 26 Lecture 10: Existential Risks & Benefits, AGI, Autonomous R&D, CBRN**

Summary: We examine AI and potential existential impacts, from CBRN (Chemical, Biological, Radiological, and Nuclear) risks to a post-scarcity future. The session also explores mechanisms by which AI could have an existential impact through autonomous research and development.

Readings:

- Bommasani, R., et al. (2024). [A path for science- and evidence-based AI policy](#). (5 pages)
- Casper, S., Krueger, D., & Hadfield-Menell, D. (2025). [Pitfalls of evidence-based AI policy](#). ICLR Blogposts Track. (13 pages)
- Gottweis, J., & Natarajan, V. (2025, February 19). [Accelerating scientific breakthroughs with an AI co-scientist](#). Google Research. (12 pages)
- Hadshar, R. (2023). [A review of the evidence for existential risk from AI via misaligned power-seeking](#). (13 pages)
- Hendrycks, D., Schmidt, E., & Wang, A. (2025). [Superintelligence strategy: Expert version](#). (34 pages)
- Morris, M. R., et al. (2024). [Levels of AGI for operationalizing progress on the path to AGI](#). International Conference on Machine Learning (ICML). (10 pages)
- Roose, K. (2025, March 14). [Powerful A.I. is coming. We're not ready](#). The New York Times. (11 pages)

- **Weds, Oct 28 Lecture 11: EU Law and Policy**

Summary: We dissect the EU AI Act and the GPAI Code of Practice, focusing on how risk classifications of AI systems map to technical characteristics like training data, model size, or capabilities. We will also explore standard-setting and conformance testing as emerging tools for ensuring technical compliance in a complex regulatory landscape. The session also considers the EU's normative goals around fundamental rights, precaution, and market harmonization.

Readings:

- Future of Life Institute. (2024). [High-level summary of the AI Act](#). (6 pages)
- Future of Life Institute. (2025). [Code of practice overview](#) [overview of the EU General-Purpose AI Code of Practice]. (41 pages)
- G'sell, F. (2024). [Regulating under uncertainty: Governance options for generative AI](#) (pp. 202–246 on the AI Act). Stanford Cyber Policy Center. (43 pages)
- Voss, A. (2024). [Getting serious about AI rules: Lack of enforcement capacity puts EU at risk](#). Euractiv. (3 pages)

- **Fri, Oct 30: Assignment due: Milestone 2: Project Plan**

Week 7: US & Global AI Governance Policy

- **Mon, Nov 2 Lecture 12: Federal & State-Level US Law and Policy**

Summary: This lecture surveys major U.S. policy actions with attention to their technical implications, such as model access restrictions and reporting requirements. We analyze the intersection between compute governance proposals and institutional capacity. Case studies include California SB1047 and SB53 thresholds for foundation model testing and the Biden Executive Order's directives on safety testing. We also discuss the fragmented nature of U.S. governance and the role of federalism, industry lobbying, and public perception.

Readings:

- California Senate Bill 53. (2025). [Transparency in Frontier Artificial Intelligence Act](#). (6 pages)
- Heim, L., & Koessler, L. (2024). [Training compute thresholds: Features and functions in AI governance](#). (28 pages)
- Hooker, S. (2024). [On the limitations of compute thresholds as a governance strategy](#). (26 pages)
- Klein, A., et al. (2024). [One year later, how has the White House AI executive order delivered on its promises?](#) Brookings Institution. (10 pages)
- The White House. (2025). [Winning the race: America's AI action plan](#). (26 pages)

- **Weds, Nov 4 Lecture 13: China and Middle Powers Approaches to AI Governance**

Summary: This lecture examines the diverse and evolving approaches to AI governance, covering China, India, and regional approaches in Southeast Asia and Africa. We situate these approaches in broader debates on technology regulation and global power asymmetries. By analyzing the institutional, legal, and normative frameworks emerging across these regions, the lecture highlights how local political economies, developmental priorities, and sociotechnical contexts shape distinct governance trajectories. It further considers the implications of these regional strategies for global governance, exploring whether they contribute to convergence toward shared norms or reinforce a more plural, fragmented landscape of AI regulation.

Readings:

- Adan, S. N., et al. (2024). [Voice and access in AI: Global AI majority participation in artificial intelligence development and governance](#). Oxford Martin AI Governance Initiative. (59 pages)
- Mohanty, A., & Sahu, S. (2024, November 21). [India's advance on AI regulation](#). Carnegie India. (40 pages)

- Munga, J., & Quansah, S. (2025, September 11). [Understanding Africa's AI governance landscape: Insights from policy, practice, and dialogue](#). Carnegie Endowment for International Peace. (14 pages)
- Personal Data Protection Commission (PDPC). (2020). [Singapore's approach to AI governance](#). (9 pages) Plus skim [AI Verify Foundation & IMDA](#). (2024). [Model AI governance framework for generative AI](#). (36 pages, skim)
- Sheehan, M. (2024, February). [Tracing the roots of China's AI regulations](#). Carnegie Endowment for International Peace. (51 pages)
- Habuka, H. (2026, October) [Japan's AI Governance - Making the Legal System AI Ready](#)
 - Must have a Stanford email to access.
- **Fri, Nov 6: Assignment due: Milestone 3: Peer Review**

Week 8: The AI Economy

- **Mon, November 9 Lecture 14: The AI Economy: The Future of Labor**

Summary: This session covers the economic dimensions of the AI transition. We use empirical research to understand how automation is reshaping labor markets, examining how AI capabilities affect different sectors and what drives substitutability or complementarity with human labor. We will explore how labor displacement intersects with inequality, worker rights, and social safety net policies.

Readings:

- Anthropic. (2025). [The Anthropic Economic Index](#).
- Boix, C., et al. (2026). [AI's economy and its political and institutional consequences](#). APSA Preprints. (42 pages)
- Brynjolfsson, E., Chandar, B., & Chen, R. (2026). [Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence](#) (updated). Stanford Digital Economy Lab. (27 pages)
- Casselman, B. (2026, July 2). [A.I. is reshaping the economy. Good luck measuring how](#). The New York Times. (2 pages)
- Murphy, C., & Brynjolfsson, E. (2026, June). [AI economic indicators: June 2026 update](#). Stanford Digital Economy Lab. (13 pages)
- **Weds, Nov 11 Lecture 15: The AI Economy: Competition, Open Models, & Antitrust**

Summary: We will survey the state of antitrust with respect to AI models, focusing on questions of liability, competition, and monopoly. Students will explore attempts to understand the emerging power of AI models within a broader legal framework.

Readings:

- Bommasani, R., et al. (2023, December). [Considerations for governing open foundation models](#). Stanford HAI Issue Brief. (10 pages)
- Hagi, A., & Wright, J. (2024). [Artificial intelligence and competition policy](#). International Journal of Industrial Organization. (16 pages)
- Lynch, S. (2026). [Open-weight models aren't enough: We need truly open-source AI models for science and society](#). Stanford HAI. (10 pages)
- Narechania, T. N., & Sitaraman, G. (2024). [An antimonopoly approach to governing artificial intelligence](#). Yale Law & Policy Review, 43, 95. (70 pages)

Week 9: Security & Human Futures

- **Mon, Nov 16 Lecture 16: Geopolitics and Global AI Governance**

Summary: We explore how technical issues like model proliferation, hardware chokepoints, and misuse risk are shaping national security strategies. Topics include the geopolitics of the semiconductor supply chain, export controls, dual-use foundation models, and proposals for international oversight of training compute. The lecture emphasizes the challenges of enforcement in a globally distributed development landscape. We also consider how AI governance intersects with power politics, norms development, differences in values across societies, and global public goods.

Readings:

- Dennis, C., et al. (2024). [What should be internationalised in AI governance?](#) Oxford Martin AI Governance Initiative. (60 pages)
- Heim, L. (2025). [Understanding the artificial intelligence diffusion framework](#). RAND Corporation. (21 pages)
- Kahl, C., & Mitre, J. (2025). [The real AI race](#). Foreign Affairs. (6 pages)
- Klein, E., & Patrick, S. (2024). [Envisioning a global regime complex to govern artificial intelligence](#). Carnegie Endowment for International Peace. (56 pages)

Supplementary Readings:

- United Nations AI Advisory Body. (2024). [Governing AI for humanity: Final report](#). United Nations. (101 pages; read at least the executive summary)

-

- **Weds, Nov 18 Lecture 17: What Does It Mean to Be Human in an AI Age?**

Summary: This lecture incorporates technical examples to probe questions of human agency, creativity, and uniqueness. Students will compare human vs. machine reasoning,

memory, and expression, and discuss whether current technical trajectories challenge our philosophical assumptions.

Readings:

- Christian, B. (2011, March). [Mind vs. machine](#). The Atlantic. (28 pages)
- Dillion, D., et al. (2023). [Can AI language models replace human participants?](#) Trends in Cognitive Sciences, 27(7), 597–600. (3 pages)
- Epstein, R. (2016). [The empty brain](#). Aeon. (16 pages)
- Griffin, L. D., et al. (2023). [Large language models respond to influence like humans](#). Proceedings of the First Workshop on Social Influence in Conversations (SICon), ACL. (9 pages)
- Harding, J., et al. (2023). [AI language models cannot replace human research participants](#). AI & Society. (2 pages)
- Koralus, P. (2025). [The philosophic turn for AI agents: Replacing centralized digital rhetoric with decentralized truth-seeking](#). arXiv. (20 pages)
- Lamparth, M., et al. (2024). [Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations](#). Conference on AI, Ethics, and Society (AIES). (9 pages)
- Leo XIV (2026). [Magnifica humanitas: On safeguarding the human person in the time of artificial intelligence](#) (encyclical letter). Vatican. (47 pages)
- Stray, J., Vendrov, I., Nixon, J., Adler, S., & Hadfield-Menell, D. (2021). [What are you optimizing for? Aligning recommender systems with human values](#). (5 pages)
- Tenenbaum, J. B., et al. (2011). [How to grow a mind: Statistics, structure, and abstraction](#). Science, 331(6022), 1279–1285. (7 pages)
- Turing, A. M. (1950). [Computing machinery and intelligence](#). Mind, 59(236), 433–460. (22 pages)

– FALL BREAK –

Week 10: Outlook & AI Futures

- **Mon, Nov 30 Lecture 18: TBD**

The content of this lecture will be decided and communicated no less than two weeks in advance.

- **Weds, Dec 2 Lecture 19: The Future of AI Governance**

Summary: This session explores a range of plausible futures for AI, including dystopian outcomes such as existential risk, utopian promises like disease eradication, and more grounded scenarios where AI progresses incrementally as a general-purpose technology. We will examine how embedded technical uncertainties in modelling ethics in code and interpretability complicate our future. One focal point will be the idea of AI as a “normal”

technology—neither apocalyptic nor revolutionary, but powerful and socially embedded—alongside contrasting visions that frame AI as fundamentally transformative or dangerous. We will reflect on the role of public imagination, risk narratives, and democratic legitimacy in shaping AI's future.

Readings):

- Bostrom, N., & Yudkowsky, E. (2014). [The ethics of artificial intelligence](#). In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. (18 pages)
- Durmus, E., et al. (2024). [Towards measuring the representation of subjective global opinions in language models](#). Conference on Language Modeling (COLM). (10 pages)
- Grayling, A. C., & Ball, B. (2024, August 6). [Philosophy has been — and should be — integral to AI](#). Social Science Space. (4 pages)
- Le Guin, U. K. (1975). [The ones who walk away from Omelas](#). *The Wind's Twelve Quarters*. (5 pages)
- Reich, R., Sahami, M., & Weinstein, J. (2021). *System error: Where big tech went wrong and how we can reboot* (chap. 8). Harper.
- Santurkar, S., et al. (2023). [Whose opinions do language models reflect?](#) International Conference on Machine Learning (ICML). (14 pages)
- Sarkar, A. (2024). [AI should challenge, not obey](#). *Communications of the ACM*, 67(10). (4 pages)
- Schaeffer, R., et al. (2023). [Are emergent abilities of large language models a mirage?](#) Conference on Neural Information Processing Systems (NeurIPS). (9 pages)
- Winner, L. (1980). [Do artifacts have politics?](#) *Daedalus*, 109(1), 121–136. (17 pages)
- **Fri, Dec 4: Assignment due: Milestone 4: Final project**