Sept 19, 2025

Governing Artificial Intelligence: Law, Policy, and Institutions

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

Monday/Wednesday 2:15 - 3:45pm
Crown Building, Stanford Law School, Room 290

Nate Persily (School of Law , Political Science, FSI, Communication)
npersily@stanford.edu
Office – Law School's Neukom Building, Room N230
Office Hours: MW 11:15-12:15
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: TBD, upon request
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 2 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.

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 designing meaningful evaluations, 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 a Senior Fellow at the Stanford Institute for HAI. 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 cochairs 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. 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 Assistant 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.

Teaching Assistants

Roberta Fischli, Course Manager and Postdoctoral Scholar in the Politics Department: fischli@stanford.edu.

Adrian Mak, Fellow, AI Initiative, Stanford Law School: ahtmak@stanford.edu.

Natalie Neufeld, Ph.D. Candidate, Department of Communication, nneufeld@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 engage the materials in discussion. ***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.*** For law students, who will be graded by Professor Persily, you must complete the group policy assignment below, but that will be graded on a pass-fail basis (and if you do not participate, you will not be eligible for Honors in the class). You must attend the lectures and may miss no more than two lectures if you wish to be eligible for Honors for the class. However, you are not expected to attend the last two classes as they conflict with the Law School's reading period. Your grade (on the H-P system and upper level class curve) will be determined by a final research paper (of no fewer than 25 pages) on any topic within the subject matter of the class. The paper is due January 6, 2026 at noon. 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.

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

Assignments and Grading Breakdown

The course includes three assignments. You will receive more information about each of the assignments well in advance of their due dates.

- Philosophy paper – due Oct 17 at 5pm PST
  - Students will write a philosophy paper of no less than 3000 and no more than 3500 words. The focus is on normative analysis. For many of you, this will be the first time you are asked to write a philosophy essay. Very detailed instructions will be provided.

- Group policy memo – due on Nov 14th at 5pm PST
  - Students will produce in groups of no more than 3 students a policy memo. A policy memo is a practical, professional document that provides specific analysis and recommendations for a specific audience. Very detailed instructions will be provided.

- Technical governance lab – due Dec 5th at 5pm PST TBC
  - Students will design an evaluation framework for a foundation model capability of their choice (e.g., deception, persuasion, fairness). The assignment involves identifying a policy-relevant risk, critiquing existing benchmarks (or other existing evaluations for that risk), and proposing an evaluation protocol with metrics, dataset design, and regulatory integration. The deliverable is a policy-technical report (8-15 pages, depending on group size) with an evaluation blueprint and critical reflection, with optional bonus credit for testing the evaluation protocol on a small set of models.

Grades will be calculated as follows:

- Technical governance lab – 30%
- Philosophy Paper – 30%
- Policy Assignment – 30%
- Attendance – 10 %

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 course readings, supplies, 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 and sections 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.

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 post-doctoral fellow 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 (max. one page). In addition, any hallucinations in submitted work will result in a failing grade for the course.

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.

Syllabus

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

- **Monday, Sep 22 Lecture 1: Introduction: Methodological Approaches & Competing Visions of the AI Transition**

  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:
  - Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025), AI 2027. AI Futures Project. (71 pages)
  - Narayanan, A., & Kapoor, S.(2025), AI as Normal Technology (Knight Foundation). (41 pages)
  - Toner, H. (2025), Unresolved Debates about the Future of AI, Rising Tide. (20 pages)

  Supplementary:
  - Dario Amodei (2024), Machines of Loving Grace. (46 pages)
  - Marc Andreessen (2023), The Techno-Optimist Manifesto (Andreessen-Horowitz). (18 pages)
  - Bengio, Y et al (2024), Managing Extreme AI Risks Amid Rapid Progress, Science. (6 pages)
  - Bengio, Y. (2025). International AI Safety Report (Executive Summary) (10 pages)

- **Wed., Sep 24 Lecture 2: History of Frontier Science & Technology Governance**

  Summary: This session examines how previous frontier technologies were governed, with a focus on the regulatory and institutional lessons that may apply to AI. We will review governance regimes for nuclear power, biotechnology, and the internet, paying particular attention to how technical characteristics of those technologies influenced oversight strategies.

  Required Readings:
  - Acemoglu, D., & Johnson, S. (2024). Prologue of *Power and Progress: Our 1000 Year Struggle Over Technology and Prosperity* (Prologue)
  - Aidinoff, M., & Kaiser, D. (2024). Novel technologies and the choices we make: Historical precedents for managing artificial intelligence. (11 pages)
  - Vermeer, M. J. D. (2024). Historical analogues that can inform AI governance. RAND Corporation. (30 pages)

- National Research Council. (1996). Understanding risk: Informing decisions in a democratic society. National Academies Press. (read pages 1-10)

Supplementary:

- Ord, T. (2022, November). Lessons from the development of the atomic bomb. (44 pages)

Week 2: Core Governance Challenges I – Technical Approaches

- **Monday, Sep 29 Lecture 3: Technical Foundations**

  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. (2024). Introduction to large language models. (Youtube video)
  - Toner, H. (2023). What are generative AI, large language models, and foundation models? Center for Security and Emerging Technology. (3 pages)
  - Wolfram, S. (2023). What is ChatGPT doing … and why does it work? (75 pages)
  - Lages, J. (2023). Direct preference optimization (DPO) - A simplified explanation. (6 pages)
  - Lee, K., et al. (2024). RLAIF vs RLHF: Scaling reinforcement learning from human feedback with AI feedback. Proceedings of the International Conference on Machine Learning (ICML). (9 pages)

- **Wednesday, Oct 1 Lecture 4: 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:

- Kleinberg, J., et al., (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Proceedings of Innovations in Theoretical Computer Science (ITCS). (17 pages)
- Kearns, M., et al., (2017). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. Proceedings of the 35th International Conference on Machine Learning (ICML). (8 pages)
- Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness & machine learning (Chapters 1, 4, and 8). MIT Press. (71 pages)
- Wang, A., et al., (2025, February). Fairness through difference awareness. (9 pages)
- Ho, D. E., & Xiang, A. (2020). Affirmative algorithms: The legal grounds for fairness as awareness. (21 pages)
- Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. (92 pages)

Week 3: Core Governance Challenges II – Political and Democratic

- **Monday Oct 6 Lecture 5: AI, Democracy, and Disinformation**

  Summary: We explore the technical mechanisms through which generative AI systems can produce misinformation and manipulate public discourse. Topics include hallucination, prompt injection, and adversarial prompting. We also discuss content moderation challenges and detection mechanisms such as watermarking and provenance tracking. The session addresses how misinformation affects electoral integrity, institutional trust, and freedom of expression.

  Readings:

  - Persily, N. (2024). Misunderstanding AI's democracy problem. Digitalist Papers.
  - Schmidt, E. (2024). Democracy 2.0.
  - Reich, O., & Calabrese, S. (2025). Beyond disinformation: How DSA risk assessments ignore democracy's real threats. Tech Policy Press. (6 pages)
  - (2024). A tech accord to combat deceptive use of AI in the 2024 elections. Munich Security Conference. (2 pages)
  - (2024). AI can help humans find common ground in democratic deliberation. Science. (8 pages)
  - (2024) How Will Advanced AI Systems Impact Democracy? (12 pages)

- **Wednesday, Oct 8 Lecture 6: Privacy**
  (Guest speakers: Yan Luo, Florence G'Sell)

Summary: This session will cover privacy, which can be implicated in the use of private data in AI model training, the use of AI to discover or disclose private information about a person, and the use of AI in surveillance.

Readings:

- Burgess, M. (2023). ChatGPT has a big privacy problem. Wired. (8 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)
- European Data Protection Board. (2024, December 17). Opinion 28/2024 on certain data protection aspects to the processing of personal data in the context of AI models. (35 pages)
- Kibby, C., & Sentinella, D. (2024). New laws in California look to the future of privacy and AI. International Association of Privacy Professionals (IAPP). (5 pages)
- Solove, Daniel (2025). Artificial Intelligence and Privacy (73 pages)

Week 4: Legal and Institutional Responses I – Domestic and Regional

- **Monday Oct 13 Lecture 7: 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:

  - European Commission. (2025). EU GPAI code of practice. (55 pages)
  - Future of Life Institute (FLI). (2024). High-level summary of the EU AI Act. (6 pages)
  - G'sell, F. (2024). Regulating under uncertainty (p. 202 to 246 on the AI Act). (43 pages)
  - Voss, A. (2024). Getting serious about AI rules: lack of enforcement capacity puts EU at risk. Euractiv. (3 pages)

- **Wednesday Oct 15 Lecture 8: 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 SB 53 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.

Special Guest: Elizabeth Kelly, former Director, U.S. AI Safety Institute

Readings:

- Hooker, S. (2024). On the limitations of compute thresholds as a governance strategy. (26 pages)
- Heim, S. (2024). Training compute thresholds: Features and functions in AI governance. (28 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)
- (2025). CA ABI 53 (6 pages)
- The White House. (2025). *America's AI action plan: Winning the race*. (26 pages)

Week 5: Legal and Institutional Responses II – Global

- **Monday Oct 20 Lecture 9: Rest of the World Approaches to AI Governance**

  Summary: This lecture examines the diverse and evolving approaches to AI governance in regions such as China, Africa, Southeast Asia, and India, situating them within 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:

  - Sheehan, M. (2024). Tracing the Roots of China's AI Regulations.
  - Mohanti, A. & Sahu, S. (2024). India's Advance on AI Regulation.
  - Adan, S. (2024). Voice and Access in AI: Global AI Majority Participation in Artificial Intelligence Development and Governance.
  - PDPC (2024). Singapore's Approach to AI Governance. + skim the model AI governance framework for generative AI.
  - Munga, J. & Quansah S. (2025). Understanding Africa's AI Governance Landscape: Insights From Policy Practice and Dialogue.

- **Wednesday Oct 22 Lecture 10: The AI Economy: Labor, Competition, and Antitrust**

Summary: This session covers the economic dimensions of the AI transition. First, 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. Second, we analyze competition and antitrust, focusing on how AI-driven scale economies, data/network effects, compute access, and foundation-model gatekeeping can entrench market power and what remedies might be needed.

Readings:

- Anthropic (2025). The Anthropic economic index.
- Kinder, M., et al. (2024). Generative AI, the American worker, and the future of work. Brookings Institution. (32 pages)
- Brynjolfsson, E., et al. (2025). Generative AI at work. NBER Working Paper. (30 pages)
- Jim VandeHei and Mike Allen (2025), A White-Collar Bloodbath, Axios. (10 pages)
- (2025). Nvidia CEO says AI will create more jobs despite workforce changes. Axios. (6 pages)
- Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). Working with AI: Measuring the occupational implications of generative AI. arXiv. (19 pages)
- Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. Stanford Digital Economy Lab. (27 pages)
- Andrei Hagiu & Julian Wright (2024), Artificial Intelligence and Competition Policy, International Journal of Industrial Organization. (77 pages)
- Tejas Narechania & Ganesh Sitaraman, An Antimonopoly Approach to Governing Artificial Intelligence, Yale Law and Policy Review, 2024. (70 pages)

Week 6: Governance Infrastructure

- **Monday Oct 27 Lecture 11: Hardware, Infrastructure and Energy**

Summary: In this session, we investigate the physical infrastructure underpinning frontier AI, including the massive energy and data center demands, and the 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:

- What Exactly are AI Companies Trying to Build?, NYTimes, 9/18/2025.
- Kneese, T., & Young, M. (2024). Carbon Emissions in the Tailpipe of Generative AI. Harvard Data Science Review. (5 pages)

- Schmidt, E. (2025). "Converting Energy into Intelligence: The Future of AI Technology, Human Discovery, and American Global Competitiveness," Testimony to House Energy and Commerce Committee. (11 pages)
- Patel, D., Nishball, D., & Eliahou Ontiveros, J. E. (2024). AI datacenter energy dilemma – Race for AI datacenter space. SemiAnalysis.
- Epoch AI. (2025). "How Much Energy Does ChatGPT Use?" Epoch AI Gradient Updates. (16 pages)
- Luccioni, S. (n.d.) The Environmental Impacts of AI - Policy Primer. (11 pages)
- Gabriel Kulp, Margaret Siu, Lennart Heim (2025) Insights from a Workshop on Hardware-Enabled Governance Mechanisms. RAND Corporation. (24 pages)

Supplementary:

- International Energy Agency, (2025). Energy and AI. (IEA). (304 pages)

- **Wednesday October 29 Lecture 12: National Security, Geopolitical Risk, and International 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.

  Required Readings:
  - Heim, L., RAND Corporation. (2025). Understanding the artificial intelligence diffusion framework. (21 pages)
  - Kahl, C., and Mitre, J. (2025). The Real AI Race, *Foreign Affairs*
    - (See Canvas if unable to access reading.)
  - Claire, D., et al. (2024). What should be internationalised in AI governance? (60 pages)
  - Klein, E., & Patrick, S. (2024). Envisioning a global regime complex to govern artificial intelligence. Carnegie Endowment for International Peace. (56 pages)
  - United Nations AI Advisory Body. (2024). Governing AI for humanity final report. United Nations. (read at least the executive summary) (101 pages)

  Supplementary:
  - RAND Corporation. (2023). Supply Chain Interdependence and Geopolitical Vulnerability: The Case of Taiwan and High-End Semiconductors, Executive Summary. (78 pages)
  - CSIS. (2025). Silicon Island: Assessing Taiwan's Importance to U.S. Economic Growth and Security. (10 pages)

- Khan, S. M. (2020). AI Chips: What They Are and Why They Matter. CSET. (72 pages)
- U.S. Department of Commerce, Bureau of Industry and Security. (2025). Department of Commerce Announces Recission of Biden-Era Artificial Intelligence Diffusion Rule, Strengthens Chip-Related Export Controls (1 page)
- Remco Zwetsloot, Helen Toner, Jeffrey Ding (2018), Beyond the AI Arms Race: America, China, and the Dangers of Zero-Sum Thinking, Foreign Affairs.

Week 7: Corporate-Level Governance

- **Monday Nov 3 Lecture 13: Organizational AI Governance**

  Summary: Governance consists of law, policy, and regulation that emanates from government. And it also involves the development of professional norms, standards, tooling, and responsibilities that emerge from the respective profession, in this case AI developers. Here we examine internal governance mechanisms such as risk assessments, model documentation (e.g., model cards), and incident response protocols. Students will evaluate how companies implement technical safeguards and ensure internal accountability for AI system behavior. The session also covers AI ethics boards and safety and security frameworks within leading labs. Finally, we will analyze corporate governance challenges, institutional incentives, and regulatory gaps in self-regulation.

  Readings:
  - Lemley, M., & Henderson, P. (2024). The mirage of artificial intelligence terms of use restrictions. (68 pages)
  - Titus, J. (2024). Scaling AI safely: Can preparedness frameworks pull their weight? Federation of American Scientists. (13 pages)
  - Anthropic. (2024). Announcing our updated responsible scaling policy. (13 pages)
  - NIST. (2023). AI Risk Management Framework (AI RMF 1.0). (34 pages)
  - Qi, X., et al. (2025). Safety Alignment Should be Made More Than Just a Few Tokens Deep. The Thirteenth International Conference on Learning Representations (ICLR). (10 pages)
  - Mökander, J., et al. (2023). Auditing large language models: A three-layered approach. Springer Nature Link. (22 pages)
  - Feffer, M., et al. (2024). Red-teaming for generative AI: Silver bullet or security theater? Association for the Advancement of Artificial Intelligence (AAAI). (16 pages)
  - Wei, A., et al. (2023). Jailbroken: How does LLM safety training fail? Conference on Neural Information Processing Systems (NeurIPS). (10 pages)
  - Reuel, A., et al. (2024). Open problems in technical AI governance. Transactions on Machine Learning Research (Chapter 1/Introduction only). (50 pages)

- **Wednesday November 5 Lecture 14: AI Evaluations As an Example of Technical AI Governance**

  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.

  Readings:
  - Longpre, S., et al. (2024). A safe harbor for AI evaluation and red teaming. International Conference on Machine Learning (ICML). (10 pages)
  - Reuel, A., et al. (2024). BetterBench: Assessing AI benchmarks. Conference on Neural Information Processing Systems (NeurIPS). (11 pages)
  - Salaudeen et al. (2025). Measurement to Meaning: A Validity-Centered Framework for AI Evaluation.
  - Boubdir, M., et al. (2023). 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. Association for Computational Linguistics (ACL). (10 pages)
  - Hadfield, G., and Clark, J. (2023). Regulatory Markets: The Future of AI Governance.arxiv. (24 pages)

Week 8: Thematic Issues

- **Monday Nov 10 Lecture 15: Data, Transparency & Access**

  Summary: Transparency – ranging from dataset documentation to broader system disclosures – will be analyzed as a central governance concern, along with broader (partially competing) values like freedom of expression, creator rights, and platform responsibility. We also discuss the technical underpinnings of access debates, including open vs. closed weights and API vs. downloadable models.

  Readings:

  - Longpre, S., et al. (2024). A large-scale audit of dataset licensing and attribution in AI. Nature Machine Intelligence. (10 pages)
  - Kaplan, J., et al. (2020). Scaling laws for neural language models. (19 pages)
  - Bommasani, R., et al. (2023). Foundation model transparency index. (63 pages, skimming is sufficient)

- Seger, E., et al. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods. LawAI Working Paper. (34 pages)
- Widder, D., et al. (2024). Why 'open' AI systems are actually closed. Nature. (5 pages)
- Solaiman, I. (2023). The gradient of generative AI release: Methods and considerations. ACM Conference on Fairness, Accountability, and Transparency (FAccT'23). (9 pages)
- Kapoor, S., et al. (2024). On the societal impact of open foundation models. International Conference on Machine Learning (ICML). (9 pages)
- Casper, S., et al. (2024). Black-box access is insufficient for rigorous AI audits. ACM Conference on Fairness, Accountability, and Transparency (FAccT'24). (9 pages)

- **Wednesday Nov 12 Lecture 16: Copyright and Free Speech**

  Summary: This lecture examines how AI model training intersects with copyright law, focusing on disputes like OpenAI vs. The New York Times. 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. In addition, AI raises new questions about defamation law and the First Amendment, in general (such as the free speech rights of robots(!), which shall be discussed.

  Readings:
  - CSR Report, June 2025, Generative Artificial Intelligence and Copyright Law (7 pages)
  - Liu et al. (2025). Language Models May Verbatim Complete Text They Were Not Explicitly Trained On. ICML. (9 pages)
  - Ippolito, D., et al. (2023). Preventing verbatim memorization in language models. Association for Computational Linguistics (ACL). (9 pages)
  - Cooper, A., et al. (2024). Machine unlearning doesn't do what you think. (24 pages)
  - Lemley, M. (2024). How generative AI turns copyright law upside down. Science and Technology Law Review, 25(2). (23 pages)
  - Cooper, A., & Grimmelmann, J. (2025). On copyright, memorization, and generative AI. Chicago-Kent Law Review, 100(1), 141–217. (76 pages)
  - Volokh, E. (2023). Large libel models? Liability for AI Output. Journal of Free Speech Law. (70 pages)
  - (2025). Denmark tries to get ahead of A.I. with copyright law for deepfakes. The New York Times. (4 pages)

Week 9: Emerging Challenges

- **Monday Nov 17 Lecture 17: The Special Problem of AI Agents & Multi-Agent Dynamics**

  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 risk from open-ended exploration. We will discuss the governance implications of decentralized, interacting agents. Guest: Tobin South

  Readings:
  - Kapoor, S., et al. (2024). AI agents that matter. (12 pages)
  - Hammond, L., et al. (2025). Multi-agent risks from advanced AI. (52 pages)
  - Kolt, N. (2024). Governing AI agents. Notre Dame Law Review, Vol. 101. (48 pages)
  - Durante, Z., et al. (2024). Agent AI: Surveying the horizons of multimodal interaction. (54 pages)
  - Wang, D., et al. (2024). LLM agent – Retrieve what you need: A mutual learning framework for open-domain question answering. Proceedings of Transactions of the Association for Computational Linguistics. (13 pages)

- **Wednesday Nov 19 Lecture 18: What Does It Mean to Be Human in an AI Age? How Should We Design for Human-Machine Companionship?**

  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:
  - Turing, A. M. (1950). Computing machinery and intelligence. (22 pages)
  - Christian, B. (2011, March). Mind vs. machine. The Atlantic. (28 pages)
  - Epstein, R. The empty brain. Aeon. (16 pages)
  - Griffin, L., et al. (2023). Large language models respond to influence like humans. Annual Meeting of the Association for Computational Linguistics (ACL) (9 pages)
  - Dillion, D., et al. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences. (3 pages)
  - Harding, J., et al. (2023). AI language models cannot replace human research participants. Springer AI and Society. (2 pages)
  - Tenenbaum, J., et al. (2023). How to grow a mind: Statistics, structure, and abstraction. Science. (7 pages)

- Lamparth, M., et al. (2023). Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. Conference on AI, Ethics, and Society (AIES). (9 pages)

- Koralus, P. (2025). The philosophic turn for AI agents: Replacing centralized digital rhetoric with decentralized truth-seeking. arXiv. (20 pages)
- Stray, Vendrov, Nixon, Adler, Hadfield-Menell, What are You Optimizing For? Aligning Recommender Systems with Human Values (5 pages)

Week 10: Synthesis and Course Wrap-Up

- **Monday Dec 1 Lecture 19: Antitrust & Defamation**

  Summary: We will survey the state of antitrust and defamation law 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 (assigned in previous weeks):
  - Andrei Hagiu & Julian Wright (2024), Artificial Intelligence and Competition Policy, International Journal of Industrial Organization. (77 pages)
  - Tejas Narechania & Ganesh Sitaraman, An Antimonopoly Approach to Governing Artificial Intelligence, Yale Law and Policy Review, 2024. (70 pages)

  - Volokh, E. (2023). Large libel models? Liability for AI Output. Journal of Free Speech Law. (70 pages) (2025).
  - Denmark tries to get ahead of A.I. with copyright law for deepfakes. The New York Times. (4 pages)
  - Ken Bensinger, Who Pays When AI Is Wrong, NY Times, Nov 12, 2025.

- **Wednesday Dec 3 Lecture 20: Ethics, Norms, and The Future of AI Governance. The Opportunities, Limits, Existential Risks, and Long-term Visions of our AI Future.**

  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 (combined from two previous sessions):

- Schaeffer, R., et al. (2023). Are emergent abilities of LLMs a mirage?, Conference on Neural Information Processing Systems (NeurIPS). (9 pages)
- Santurkar, S., et al. (2023). Whose opinions do language models reflect?, International Conference on Machine Learning (ICML). (14 pages)
- Bostrom, N., & Yudkowsky, E. The ethics of artificial intelligence. Cambridge University Press. (18 pages)
- Winner, L. (1980). Do artifacts have politics? (17 pages)
- Grayling, A., & Ball, B. (2024, August 6). Philosophy has been – and should be – integral to AI. The Conversation. (4 pages)
- Durmus, E., et al. (2024). Towards measuring the representation of subjective global opinions in language models., COLM (10 pages)
- Sarkar, A. (2025). AI should challenge, not obey. arXiv. (4 pages)
- Reich, R., Sahami, M., and Weinstein, J. (2021). Chapter 8 of *System Error: Where Big Tech Went Wrong and How We Can Reboot*.
- leGuin, Ursula. The Ones Who Walk Away From Omelas.
- Roose, K. (2025). Powerful A.I. Is Coming. We're Not Ready. The New York Times. (11 pages)
- Morris, M., et al. (2024). Levels of AGI for operationalizing progress on the path to AGI. International Conference on Machine Learning (ICML). (10 pages)
- Hadshar, R. (2023). A review of the evidence for existential risk from AI. (13 pages)
- Bommasani, R., et al. (2024). A path for science- and evidence-based AI policy. (5 pages)
- Casper, S., et al. (2025). Pitfalls of evidence-based AI policy. (13 pages)
- Gottweis, J., & Natarajan, V. (2025, February 19). Accelerating scientific breakthroughs with an AI co-scientist. Google Research. (12 pages)
- Dan Hendrycks, Eric Schmidt and Alexandr Wang (2025) Super-Intelligence Strategy: Expert Version. (34 pages)