

This syllabus is pretty old and in my opinion not that good—[this newer syllabus](#) is probably a better use of your time.

Sources of Readings

This syllabus is a compilation of several older reading lists:

1. [Ashwin recommendations](#)
2. [Earlier reading group recommendations](#), from Oxford
3. [Felipe recommendations](#)
4. Week 5 of this [social sciences and existential risk syllabus](#)

1. Intro to AI Governance

What does the field of AI Governance focus on? What major concerns, perspectives, and goals guide the field?

Core materials:

- [AI Governance: Opportunity and Theory of Impact](#) - Dafoe
- Pp. 5-13 of [AI Governance: A Research Agenda](#) - Dafoe

Further reading:

- [80,000 Hours problem profile on positively shaping the future of AI](#)
 - See [here](#) for recommendations on potential careers in this space
- Wait But Why's The Artificial Intelligence Revolution [Parts 1 and 2](#)
- [Response post](#) to Wait but Why
- [“The Big Picture”](#) - 80,000 Hours (section of an article)
 - A refresher on some key concepts around pursuing high-impact careers

2. Forecasting

When, if ever, will AI transform society as drastically as the industrial revolution did? When, if ever, will AI capabilities in a wide range of tasks—including complex tasks about which there is little data—far exceed those of humans? Researchers have taken varied approaches to

forecasting AI progress: surveying AI experts, extrapolating trends in compute and algorithms, using “semi-informative priors,” and extrapolating macroeconomic trends.

Core Materials

- [When Will AI Exceed Human Performance? - Grace et al.](#)
 - (Feel free to skim, focusing on the graphs and tables)
 - This paper analyzes the results of a survey of machine learning researchers.
- [AI and Compute - Amodei et al.](#)
 - This post estimates that the computing power used by leading machine learning models has been doubling every 3.4 months from 2012 to 2018. It argues that this is a reason to expect AI capabilities to drastically improve soon.
- [Reinterpreting “AI and Compute” - Garfinkel](#)
 - A response to the previous post, arguing that the same trend implies opposite conclusions.

Further Reading

- Broad discussions of forecasting AI:
 - [What Do We Know about AI Timelines?](#) - Open Philanthropy
 - [Danny Hernandez on Forecasting and the Drivers of AI Progress](#) - 80,000 Hours Podcast
 - [How well can we actually predict the future? Katja Grace on Why Expert Opinion isn't a Great Guide to AI's Impact and How to Do Better](#) - 80,000 Hours Podcast
 - For related research, see [aiimpacts.org](#)
- More on expert surveys:
 - [When Will AI Exceed Human Performance? Evidence from AI Experts](#) - Müller and Bostrom
 - [How Feasible Is Long-range Forecasting?](#) - OpenPhil
- More on extrapolation of compute trends:
 - [Draft report on AI timelines - Cotra, summary by Shah](#)
 - See also [AI and Efficiency](#) which shows that the compute needed to train neural nets has been halving every 16 months.
 - [Interpreting AI Compute Trends](#) - Carey
 - ["AI and Compute" trend isn't predictive of what is happening](#) - Zhov
- Semi-informative priors:
 - [Report on Semi-informative Priors](#) - Davidson
- Macroeconomic forecasting (often long and technical):
 - [“Does Economic History Point Toward a Singularity?”](#) - Garfinkel (2020)
 - [“Modeling the Human Trajectory”](#) - Roodman (2020)

- [“Artificial intelligence and economic growth”](#) - Aghion et al. (2017)
- [“Are we approaching the economic singularity?”](#) - Nordhouse (2015)

3. AI and China

Core Materials

[China’s Current Capabilities, Policies, and Industrial Ecosystem in AI. Jeffrey Ding, GovAI](#)

This testimony compares the current AI capabilities of China and the U.S. by slicing up the fuzzy concept of “national AI capabilities” into three cross-sections: 1) scientific and technological inputs and outputs, 2) different layers of the AI value chain and 3) different subdomains of AI. This approach reveals that China is not poised to overtake the U.S. in the technology domain of AI. The testimony makes policy recommendations to maintain the status quo via reviving the Office of Technology Assessment, building bridges across the “Valley of Death” in the AI domain, and increasing attention to the risks of accidents and emergent effects associated with the deployment of emerging technologies related to AI.

[Understanding China’s AI Strategy: Clues to Chinese Strategic Thinking on Artificial Intelligence and National Security, Gregory Allen, Centre for a New American Security](#)

Meeting-informed judgments about Chinese leadership’s views, strategies, and prospects for AI’s applications to China’s economy and national security.

Further Reading

Military

- [National Security Commission on AI: Interim Report](#)
- [The Logic of Strategic Assets : From Oil to AI](#) (2020) - by Jeff Ding and Allan Dafoe
- [Innovation and National Security: Keeping our Edge](#) by CFR
- Gilli & Gilli, 2019. Imitation, innovation, disruption: Challenges to NATO superiority in military technology. NDC Policy Brief
- Horowitz: [AI, International Competition, and the Balance of Power](#)
- [Work and Grant, 2019](#). Beating the Americans at their Own Game. An Offset Strategy with Chinese Characteristics.

“Structural”

- [Goldsmith, J., Russell, S.](#) **Strengths Become Vulnerabilities: How A Digital World Disadvantages the US in its International Relations.** A Hoover Institution Essay

- [Weaponized Interdependence](#)
- [Four Internets: The Geopolitics of Digital Governance](#)

Cybersecurity

- Nye, J. (2019). Protecting Democracy in an Era of Cyber Information War (Belfer Center)

Chinese AI

- [Deciphering China's AI Dream](#)
- [The Chinese Social Credit System: A Model for Other Countries?](#) (38 pp.)
- [Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure](#) (39 pp.)

Semiconductors

- [Maintaining the AI Chip Competitive Advantage of the United States and its Allies](#) (CSET)
- [Recommendations on Export Controls for Artificial Intelligence](#) (CSET)

Safety cooperation

- [AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement](#) (CSET report)
- [Why Responsible AI Development Needs Cooperation on Safety](#) (OpenAI)

Near term vs long term

Core Materials

[Bridging near- and long-term concerns about AI, Stephen Cave, Seán ÓhÉigeartaigh, Leverhulme Centre for the Future of Intelligence](#)

A succinct case for breaking down the dichotomy that is often drawn between near-term versus long-term concerns about AI.

[Near term versus long term AI risk framings by Carina Prunkl \(FHI\) and Jess Whittlestone, Leverhulme Centre for the Future of Intelligence](#)

This article considers the extent to which there is a tension between focusing on the near and long-term AI risks.

Further Reading

[AI Alignment Podcast: On the Long-term Importance of Current AI Policy with Nicolas Moës and Jared Brown](#)

In this podcast discussion, two policy experts propose several reasons why people concerned with long-term risks from AI should work on near-term policy problems.

[Europe proposes strict A.I. regulation likely to have an impact around the world](#)

This recent (as of 2021) EU proposal on AI regulation is both notable in its own right, and it arguably serves as an example of how laws that may create lasting path-dependencies are being developed now.

US Public Opinion on AI

US public opinion on AI is one factor that influences the political feasibility of various policy proposals for governing AI. What do Americans currently think about the US government's role in regulating AI?

[US Public Opinion on Artificial Intelligence, Baobao Zhang, Allan Dafoe, GovAI](#)

Results of a survey of 2,000 Americans, covering varied topics, and examining historical and cross-national trends in public opinion regarding AI.

[Public opinion lessons for AI regulation, Baobao Zhang, GovAI](#)

This brief focuses on how public opinion will likely shape the regulation of three applications of AI in the U.S.: (1) facial recognition technology used by law enforcement, (2) algorithms used by social media platforms, and (3) lethal autonomous weapons.

[remainder is in flux]

Week 2: AI Through Short-Term and Long-Term Perspectives

More Shortermist Readings

- **Is This Time Different? The Opportunities and Challenges of Artificial Intelligence** (skimming the headlines should be enough):

https://obamawhitehouse.archives.gov/sites/default/files/page/files/20160707_cea_ai_furman.pdf

- **The impact of artificial intelligence on human society and bioethics** (especially check out "Negative/Positive impacts" and "Artificial intelligence ethics must be developed"): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7605294/>

More Long Termist readings

- **AI Governance: Opportunity and Theory of Impact** by Allan Dafoe
<https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact>
- **Concrete Problems in AI Safety** (skimming through headlines and bullet points should be enough) <https://arxiv.org/abs/1606.06565>

Optional

- **Longtermist**
 - **Ben Garfinkel on scrutinizing classic AI risk arguments**
<https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/>
 - What Failure Looks Like:
<https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like>
- **Shorttermist**
 - AI Blackbox:
<https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>
 - ML & Healthcare
<https://emerj.com/ai-sector-overviews/machine-learning-healthcare-applications/>
& <https://fortune.com/2016/06/29/ibm-watson-cancer-moonshot/>

Helpful links

- 80K's recommendations
 - <https://80000hours.org/articles/us-ai-policy/#further-reading> - not what we are looking for
 - <https://80000hours.org/articles/ai-policy-guide/#resources>
 - <https://docs.google.com/document/d/1W7KmO7TcHbM1-0CFW7TrbvG8PFqEcoYeie-vv994d6E/edit#heading=h.64woebv06g86>
- https://docs.google.com/document/d/1pre7zl03nVmAX3KA_S5OCqJC9op99rDxWOpGdcfeoYo/edit (VERY interesting)
- Oxford Spring 2020
 - https://docs.google.com/document/d/18lUsH6o7ZsknEHYbYZzGD_9ZwWoBrFuK-Xdzs8aQpQg/edit#
 - <https://forum.effectivealtruism.org/posts/eLKX9bmra9ZR2AQzD/ai-governance-reading-group-guide>
- Governance of AI: from Ashwin and a bunch of people at FHI https://docs.google.com/document/d/1M1MEjmNG1tf3XkNSiD70hXfFJ1KyN_AUjMhJ27bSs_I/edit#heading=h.10qr9okbfzw5
- From Mauricio / Felipe
 - Social Sciences and X risk reading group:
<https://docs.google.com/document/d/1hdvJLNL8vI7rGPPJHGrJme8LwvJWYz01o3yEVDdym2s/edit#>
 - Felipe's resources:
https://docs.google.com/document/d/1_7ey8eXkaFoRla9wjVA0wNaRYzDKR340lBS2kJ5qW6s/edit
- Iterated Distillation and Amplification
 - Summary of Paul Christiano's Iterated Amplification AI safety agenda
<https://docs.google.com/document/d/11QGpURtFF-JFnWjkdlybW9Q-ml3fiKJjhxc0-LjprNQ/edit#>
 - Machine Learning Projects for Iterated Distillation and Amplification
https://owainevans.github.io/pdfs/evans_ida_projects.pdf
 - <https://80000hours.org/podcast/episodes/paul-christiano-ai-alignment-solutions/#ida>

