AI Policy Through An EA Lens - Readings

This is a living document. Feel free to add suggestions/thoughts / etc. (Very) tentative readings for the upcoming week can be found by scrolling through the rest of the document.

Week 1: What does AI Policy look like from an EA perspective? What can we learn from people outside of Effective Altruism?

- For a refresher on EA concepts: https://80000hours.org/key-ideas/
- 80,000 Hours problem profile on positively shaping the future of AI
- Optional (Very Fun) Readings:
    - Wait But Why's The Artificial Intelligence Revolution Parts 1 and 2
    - Response post to Wait but Why

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
- **Shortermist**
    - 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/18IUsH6o7ZsknEHYbYZzGD_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 FHIhttps://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_7ey8eXkaFoRla9wjVA0wNaRYzDKR340lBS2kJ5gW6s/edit
- Iterated Distillation and Amplification
  - Summary of Paul Christiano's Iterated Amplification AI safety agenda https://docs.google.com/document/d/11QGpURtFF-JFnWjkdIybW9Q-mI3fiKJjhxc0-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

Week Planning?

- *** **AI Governance: Opportunity and Theory of Impact by Allan Dafoe**
- **https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments/**
    - Transcript
    - Rob's intro [00:00:00]
    - The interview begins [00:03:20]
    - Long-run implications of AI [00:06:48]
        - Industrial revolution - How will the industrial revolution go well? How do you know where to start?
        - Electricity
        - Agriculture
        - Things that are super far oof → more neglected, and maybe have more influence?
        - Biorisk -> ethics of genetic engineering
    - Political instability [00:15:25]
        - Less deference possibly due to better targetting
    - Lock-in scenarios [00:23:01]
        - US constitution
        - The US being dominant power -> democracy
        - Egypt is not dominant power anymore
        - Neolithic revolution (hunter-gather to agricultural)
            - General roles
            - Autocracy rather than collective decisions
            - Disease
        - Norms on the moral status
    - Alignment problem [00:33:31]
        - Why Bostrom / Yudkowsky? More prominent, a whole book on this and many essays
    - A high-level overview of Bostrom [00:39:34]
        - Superintelligence
        - Paperclip doesn't seem concrete
        - People don't build bridges that fall / there will be safety measures
    - Brain in a box [00:50:11]
        - The jump to the AI being radically smarter
        - Specialization happens vs. superintelligent AI
    - Contenders for types of advanced systems [00:58:42]
        - There are things that AI can do that human can't do like Deep Fakes
        - How do we know what humans can do that AI can't do
    - Implications of smoothness [01:04:07] (I think this means smooth transition?)
        - Humans and chimpanzees aren't that different but humans run the world; sudden discontinuity
        - Discontinuity for classic arguments (10%?) vs. smooth transition
    - Groups of people who disagree [01:18:28]
        - Classic (Bostrom / Yudkowsky)

        - We need more compute
    - Orthogonality thesis [01:25:04]
        - If it were so smart then it'd know that humans didn't want the AI to do that
        - Inverse reinforcement learning to infer goals
    - Entanglement and capabilities [01:37:26]
    - Instrumental convergence [01:43:11]
        - Accruing power/instruments to, for instance, make paperclips
        - Bad to try to predict based on the possible ways to make the technology since there are just too many possible ways
        - Process > possible ways
        - mesa-optimization
    - Treacherous turn [02:07:55]
        - Seems like Ben less thinks that AI being
        - There's a big multiplier if you're working on an area that you find intrinsically fascinating, or that you feel like your natural talent or skills are well set up to do
            - I would probably be uncomfortable given the current state of arguments of encouraging someone to switch from the area where they feel a more natural inclination to work on AI stuff.
    - Where does this leave us? [02:15:19]
    - What role should AI play in the EA portfolio? [02:22:30]
        - There's a big multiplier if you're working on an area that you find intrinsically fascinating, or that you feel like your natural talent or skills are well set up to do
            - I would probably be uncomfortable given the current state of arguments of encouraging someone to switch from the area where they feel a more natural inclination to work on AI stuff.
        - Getting started
            - Get familiar with basic economics concepts
            - Make blog posts
        -
    - Rob's outro [02:37:51]
- https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like
    - Part I: You can't rely on trial and error and easy to measure outcomes
    - Part II: take intermediate steps (gaining influence) to better reach their goals
- ~~https://www.lesswrong.com/posts/Z5ZBPEgufmDsm7LAv/what-can-the-principal-agent-literature-tell-us-about-ai~~
    - ~~I forgot what rents were~~
- **Optional**
    - ***** Why responsible AI development needs cooperation on safety by OpenAI**
        - I did like this article

Week 3 - Working on the Field

- *** The case for building expertise to work on US AI policy by Niel Bowerman (kind of US-centric but probably good // revise)
- Personal thoughts on careers in AI policy and strategy by Carrick Flynn

Week 4

- * When will AI exceed human performance by Grace, et al.
- Takeoff speeds by Paul Christiano

Week 5 - Case studies (choose a country from the list)

- https://docs.google.com/document/d/1pre7zl03nVmAX3KA_S5OCqJC9op99rDxWOpGdcfeoYo/edit

Week 6 - AI in Society

- https://www.youtube.com/watch?v=fMym_BKWQzk&t
- https://nickbostrom.com/ethics/artificial-intelligence.pdf

Contents

*Note that the syllabus below is from **EA Oxford.***

Contents

Intro to the field

Forecasting

Near term vs long term

AI and China

Openness and publication norms

AI's effect on warfare

Public opinion on AI

Towards Cooperation

Concrete Policy Recommendations

1. Intro to the field

AI Governance: A Research Agenda, Allan Dafoe, GovAI

GovAI's research agenda provides a broad view of the AI governance research landscape – themes, open questions, and pathways for future work in this space.

Some Background on Our Views Regarding Advanced Artificial Intelligence, Holden Karnofsky, Open Phil

The Open Philanthropy Project is a significant funder of AI strategy research. This post from their CEO, Holden, summarizes the case for why advanced AI warrants substantial attention as one of the most significant sources of risk for humanity. In particular, this post provides a clear conceptual definition of Transformative AI (TAI) in relation to other commonly used terms such as artificial general intelligence (AGI) and superintelligence.

Thinking About Risks From AI: Accidents, Misuse and Structure, Remco Zwetsloot, Allan Dafoe, GovAI

A short, conceptual piece articulating the three categories of risk that arise from advanced AI - malicious use risk, accident risk, and structural risk.

Further reading:

The Precipice: Existential Risk and the Future of Humanity, Toby Ord, FHI

How our long-term future could be better than almost anyone believes, but also how humanity's recklessness is putting that future at grave risk. Worth reading even if you know a fair bit about existential risk already.

Human Compatible: Artificial Intelligence and the Problem of Control, Stuart Russel, CHAI

Russell begins by exploring the idea of intelligence in humans and in machines. He describes the near-term benefits we can expect and outlines the AI breakthroughs that still have to happen before we reach superhuman AI. He also spells out the ways humans are already finding to misuse AI.

Russell suggests that we can rebuild AI on a new foundation, according to which machines are designed to be inherently uncertain about the human preferences they are required to satisfy. Such machines would be humble, altruistic, and committed to pursuing our objectives, not theirs. This new foundation would allow us to create machines that are probably deferential and probably beneficial.

Superintelligence: Paths, Dangers, Strategies, Nick Bostrom, FHI

2014 book arguing that if machine brains surpass human brains in general intelligence, then this new superintelligence could replace humans as the dominant lifeform on Earth. Sufficiently intelligent machines could improve their own capabilities faster than human-computer scientists, and the outcome could be an existential catastrophe for humans.

2. Forecasting

When Will AI Exceed Human Performance? Katja Grace, John Salvatier (Department of Political Science, Yale University), Allan Dafoe, Baobao Zhang, Owain Evans, GovAI
This paper uses survey data from machine learning researchers to predict when AI will exceed human performance in a variety of domains. For additional research on AI progress timelines, see aiimpacts.org

AI and Compute, Dario Amodei, Danny Hernandez, Open AI
This post shows that between 2012 and 2018 the computing power used by the most computationally intensive machine learning projects has doubled every 3.4 months. It argues improvements in compute have been a key component of AI progress, and that if this trend continues, we should expect to see systems far outside today's capabilities soon.
See also AI and Efficiency which shows that the compute needed to train neural nets has been halving every 16 months.

Reinterpreting "AI and Compute", Ben Garfinkel. guest post for AI Impacts
A response to the previous post, explaining why the increases in compute usage might actually be evidence against dramatically more capable systems coming soon

Further reading:
Forecasting the Drivers of AI Progress - *80,000 Hours Podcast with Danny Hernandez*
How well can we actually predict the future? - *80,000 Hours Podcast with Katja Grace*

3. Near term vs long term

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.

4. AI and China

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

From meetings with senior Chinese officials and corporate executives at diplomatic, military, and private-sector AI conferences Gregory has arrived at a number of key judgments about Chinese leadership's views, strategies, and prospects for AI as it applies to China's economy and national security.

5. Openness and publication norms

Strategic Implications of Openness in AI Development, Nick Bostrom, FHI

Bostrom gives considerations around different forms of openness in AI development. An excellent example of analyzing how a specific strategic variable affects the development of advanced artificial intelligence. This provides a basis upon which we can analyze the impact of proposed policies and norms in AI development on openness (e.g. intellectual property regimes, AI export controls).

The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse of the Technology? Toby Shevlane, Allan Dafoe, GovAI

The existing conversation within AI has imported concepts and conclusions from prior debates within computer security over the disclosure of software vulnerabilities. While disclosure of software vulnerabilities often favors defense, this article argues that the same cannot be assumed for AI research. It provides a theoretical framework for thinking about the offense-defense balance of scientific knowledge.

6. AI's effect on warfare

How Might Artificial Intelligence Affect the Risk of Nuclear War? Edward Geist, Andrew Lohn, RAND

This piece places the intersection of AI and nuclear war in a historical context and characterizes the range of expert opinions. It then describes the types of anticipated concerns and benefits through two illustrative examples:

1) *AI for detection and for tracking and targeting and*
2) *AI as a trusted adviser in escalation decisions*

Even with modest rates of technological progress, AI has the potential to exacerbate emerging challenges to nuclear strategic stability by the year 2040

Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority, Richard Danzig, Centre for a New American Security

This report provides a clear, compelling argument for how advances in military technology (including but not limited to AI) can impede relevant decision making and create risk, thus demanding greater attention by the national security establishment.

Further reading:

How Does the Offense-Defense Balance Scale? Ben Garfinkel, Allan Dafoe, GovAI

The article asks how the offense-defense balance scales, meaning how it changes as investments into a conflict increase. Simple models of ground invasions and cyberattacks that exploit software vulnerabilities suggest that, in both cases, growth in investments will favor offense when investment levels are sufficiently low and favor defense when they are sufficiently high. We refer to this phenomenon as offensive-then-defensive scaling or OD-scaling. Such scaling effects may help us understand the security implications of applications of artificial intelligence that in essence scale up existing capabilities.

When speed kills: Lethal autonomous weapon systems, deterrence, and stability

While the applications of AI for militaries are broad, lethal autonomous weapon systems (LAWS) represent one possible usage of narrow AI by militaries. Research and development on LAWS make exploring the consequences for security a crucial task. This article draws on classic research in security studies and examples from military history to assess the potential development and deployment of LAWS, as well as how they could influence arms races, the stability of deterrence, including strategic stability, the risk of crisis instability, and wartime escalation.

The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk

This edited volume focuses on the impact of AI on nuclear strategy. It assembles the views of 14 experts from the Euro-Atlantic community on why and how machine learning and autonomy might become the focus of an armed race among nuclear-armed states; and how the adoption of these technologies might impact their calculation of strategic stability and nuclear risk at the regional level and trans-regional level.

7. Public opinion on AI

US Public Opinion on Artificial Intelligence, Baobao Zhang, Allan Dafoe, GovAI

In the report, we present the results from an extensive look at the American public's attitudes toward AI and AI governance. We surveyed 2,000 Americans with the help of YouGov. As the study of public opinion toward AI is relatively new, we aimed for breadth over depth, with our questions touching on workplace automation; attitudes regarding international cooperation; the public's trust in various actors to develop and regulate AI; views about the importance and likely impact of different AI governance challenges; and historical and cross-national trends in public opinion regarding AI.

Public opinion lessons for AI regulation, Baobao Zhang, GovAI

84% of the American public believes that AI is a technology that should be carefully managed. 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.*

These case studies were selected because they involve AI governance issues that the American public characterize as either highly likely to impact them in the next decade or important for tech companies and governments to manage.

8. Towards Cooperation

(Each of these readings is long, so just make sure that at least one person reads each piece.)

Why Responsible AI Development Needs Cooperation on Safety, Amanda Askell, Miles Brundage, Jack Clark, Open AI

This paper identifies four strategies that can be used today to improve the likelihood of long-term industry cooperation on safety norms in AI:

1) *communicating risks and benefits,*
2) *technical collaboration,*
3) *increased transparency, and*
4) *incentivizing standards.*

The analysis shows that industry cooperation on safety will be instrumental in ensuring that AI systems are safe and beneficial, but competitive pressures could lead to a collective action problem, potentially causing AI companies to under-invest in safety.

The Windfall Clause, Cullen O'Keefe, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, Allan Dafoe, GovAI

Over the long run, technology has improved the human condition. Nevertheless, the economic progress from technological innovation has not arrived equitably or smoothly. While innovation often produces great wealth, it has also often been disruptive to labor, society, and world order. In light of ongoing advances in artificial intelligence ("AI"), we should prepare for the possibility of extreme disruption, and act to mitigate its negative impacts. This report introduces a new policy lever to this discussion: the Windfall Clause.

Agile Alliances, Andrew Imbrie, Ryan Fedasiuk, Catherine Aiken, Tarun Chhabra, Husanjot Chaha, CSET

This report recommends a three-pronged strategy to strengthen U.S. alliances and partnerships in A:

1) *defend against the threats posed by digital authoritarianism,*
2) *network with like-minded countries to pool resources and accelerate technological progress, and*
3) *project influence and leverage safe and reliable AI in support of inclusive growth, human rights, and liberal democratic values.*

The report goes on to recommend 10 concrete initiatives with specific partners to achieve these aims.

9. Concrete Policy Recommendations

Keeping Top AI Talent in the United States, Remco Zwetsloot, CSET

More than half of the AI workforce in the United States was born abroad, as were around two-thirds of current graduate students in AI-related fields. Tens of thousands of international students get AI-related degrees at U.S. universities every year. Retaining them, and ensuring a steady future talent inflow, is among the most important things the United States can do to address persistent domestic AI workforce shortages and to remain the global leader in AI. This report explains the mechanisms of student retention, outlines the policy context, and makes suggestions for US policymakers.

Recommendations on Export Controls for Artificial Intelligence, Carrick Flynn, CSET

The U.S. government has recently begun to pay close attention to the export of artificial intelligence and AI-relevant technologies. Export controls are complicated policy tools that require the careful balancing of competing interests and priorities. This paper clarifies the stakes by reviewing and assessing options for export controls on AI software, algorithms, data sets, chips, and chip manufacturing equipment. The key takeaway is that chip manufacturing equipment is likely to be a highly effective point of export control, with other areas likely to be either ineffectual or affirmatively damaging to the interests of the United States and its democratic allies.