AI policy and security/safety introductory reading list

This reading list is focused on AI policy from a security/safety perspective. For reading lists from other perspectives see, for example, here, here or here.

*** stars represent how high of a priority I give to a piece of content. To add/remove recommendations, please email niel@80000hours.org

Introduction

- *** 80,000 Hours problem profile on positively shaping the future of AI
- More informal and less dense alternative: Wait But Why Parts 1 and 2

Newsletters

- *** ImportAI by Jack Clark
- *** Policy.ai by CSET
- AI alignment newsletter by Rohin Shah
- The Algorithm by MIT
- ChinAI newsletter by Jeff Ding
- AI.Westminster by Elliot Jones
- EuropeanAI newsletter by Charlotte Stix
- German language AI newsletter by Stiftung NV

AI policy

- *** AI Governance: Opportunity and Theory of Impact by Allan Dafoe
- *** Why responsible AI development needs cooperation on safety by OpenAI
- ** NSCAI final report
- * Killer Apps by Paul Scharre
- * CSET's policy reports series
- * Malicious use of AI
- * AI Governance: A Research Agenda by Allan Dafoe
- * Deciphering China's AI dream by Jeff Ding
- Understanding China's AI strategy by Greg Allen
- How might AI affect the risk of nuclear war by RAND
- Battlefield Singularity by Elsa Kania
- Politics + AI readling list by Tim Dutton
- AI governance career paths for Europeans by Stefan Torges

AI safety

- *** Human Compatible by Prof. Stuart Russell
- ** Concrete problems in AI safety [summary, podcast]
- * Building safe AI by the DeepMind safety team
- * AI alignment research overview by Jacob Steinhardt
- * Chris Olah's views on AGI safety
- Buck's AI safety reading list

AI technology trajectory speculation

- * When will AI exceed human performance by Grace, et al.
- * The Precipice chapter on AI, by Toby Ord
- * AI and compute by OpenAI
    - Reinterpreting AI and compute by Ben Garfinkel
- Takeoff speeds by Paul Christiano
    - Likelihood of discontinuous progress around the development of AGI by AI Impacts
    - I haven't seen good write-ups in response to these two articles, but I found the discussion in the comments here useful. To dive deeper see this reading list.

Working in the field

- *** The case for building expertise to work on US AI policy by Niel Bowerman
- ** Preparing for federal jobs
- * Guide to working in AI policy and strategy by Miles Brundage (this is now somewhat out of date, but still valuable)
- Personal thoughts on careers in AI policy and strategy by Carrick Flynn
- Webinar on getting hired into policy roles by Georgetown School of Foreign Service Careers Service

Understanding machine learning

- * AI reading list by Vishal Maini

Podcasts

- 80,000 Hours podcasts
    - ** Allan Dafoe
    - ** Helen Toner
    - ** Brain Christian
    - * Ben Garfinkel on scrutinising classic AI risk arguments
    - * Miles Brundage
    - * Paul Christiano: 1 and 2

    - * Tom Kalil
    - * Jeff Ding
    - Katja Grace
    - OpenAI
- Cyberlaw podcast
    - Miles Brundage and Shahar Avin
    - Michael Page
- CNAS
    - Helen Toner and Jack Clark
    - Strategic Competition in the age of AI
- Y-Combinator podcast
    - Miles Brundage and Tim Hwang

Meta-learning

- * Michael Nielsen on how to learn new fields with Anki

Further reading

- * Coursera course on Neural Networks and Deep Learning by Ng, et al.
- Global Politics of AI reading list by Dafoe, et al.
- List of books about disasters, near-misses, and the people who prevented them
- AI and International Security Syllabus by Remco Zwetsloot
- CNAS AI and Global Security reading list
- China and AI reading list by Jeff Ding
- Center on Human-compatible AI bibliography for lots of technical AI safety reading