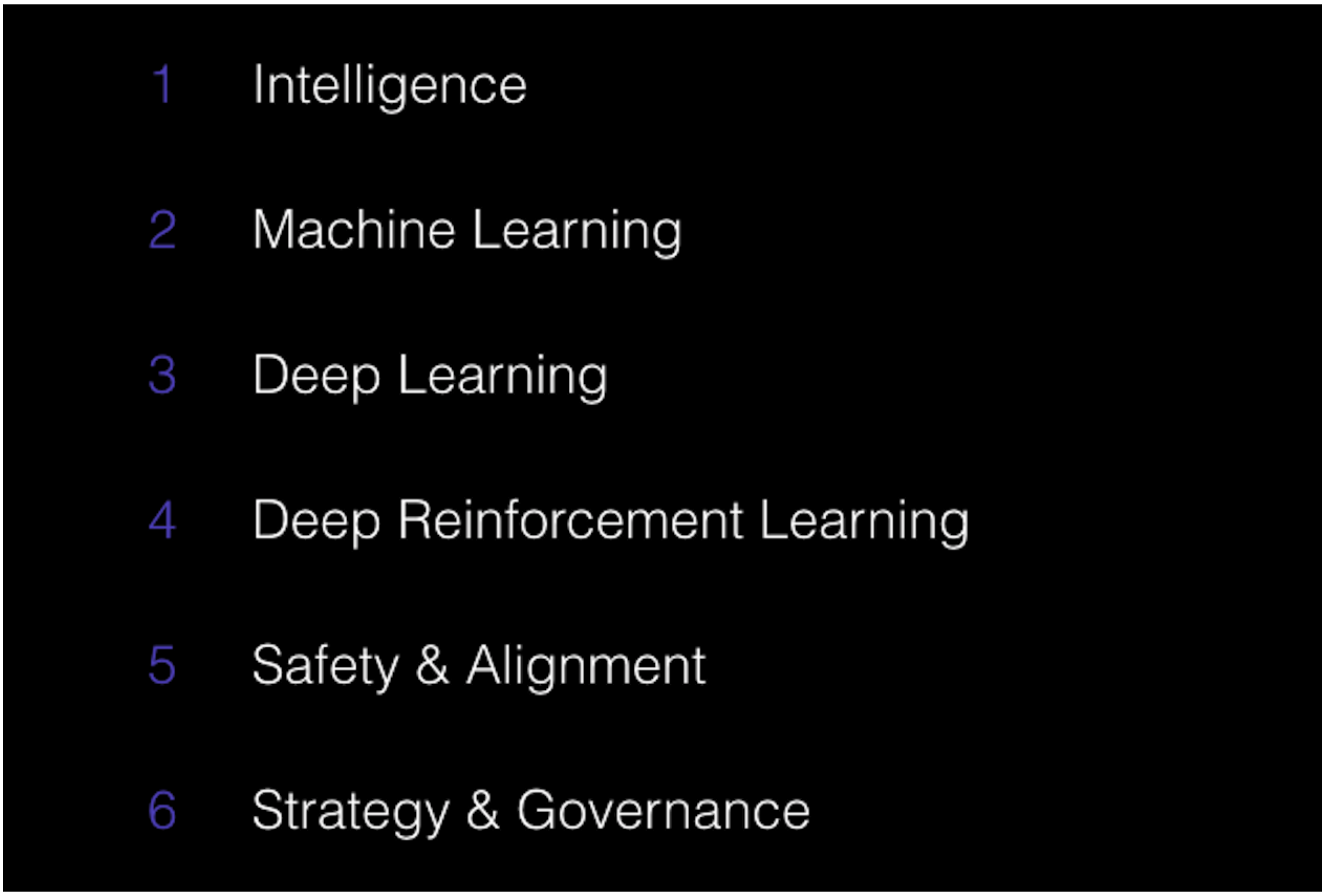
AI Reading List

Vishal Maini
Jun 19 · 5 min read

For newcomers to the field of artificial intelligence, prioritizing among endless AI resources can be an overwhelming challenge. This list attempts to do exactly that: it's a carefully curated compilation of resources for getting up to speed quickly on key topics in artificial intelligence research and its long-term implications.

The list is divided into "80/20" sections with a few high-priority readings, for maximum value with minimal time investment, and "deep dive" sections for further exploration.

Readers need not be technical, nor have a prior background in artificial intelligence. The list may be of special interest to those considering entering the field of AI research or adjacent fields, whether in technical or non-technical roles.



1 Intelligence

2 Machine Learning

3 Deep Learning

4 Deep Reinforcement Learning

5 Safety & Alignment

6 Strategy & Governance

This list covers high-level context (”What is intelligence, and what would it mean to re-create it in machines?”) technical foundations (”How does narrow AI work today, and what are some of the favored technical approaches towards general AI?”), safety considerations (”What will it take for beyond-human-level AI to be safe and act in accordance with human preferences?”), and strategic questions (”How can we coordinate towards beneficial outcomes from advanced AI?”).

```
To get the most value out of these resources, consider actively
reading. Write short summaries of key concepts and takeaways that
you can refer to later (see: Feynman technique). Spend some time
critically analyzing the ideas. What makes intuitive sense to you,
and what doesn’t? What are your criticisms of the ideas presented?
How do concepts fit together?
```

1. Intelligence

What is intelligence, and how might we re-create it in machines? Why now? Three ingredients for AI progress — compute, data, and algorithms.

80/20

- The AI Revolution: The Road to Superintelligence — Tim Urban (Wait But Why), 2015
- Superintelligence: Paths, Dangers, Strategies, Ch. 1–3 — Nick Bostrom, 2014
- The Power of Self-Learning Systems (video) — Demis Hassabis, 2019
- How to get empowered, not overpowered, by AI (video) — Max Tegmark, 2018
- Can we rule out near-term AGI? (video) — Greg Brockman, 2018
- AI and Compute — OpenAI, 2019

Deep dive

- Computing machinery and intelligence — Alan Turing, 1950
- Machine Super Intelligence — Shane Legg, 2008
- Reframing Superintelligence — Eric Drexler, 2019
- AlphaGo (documentary) — DeepMind, 2017
- The Measure of All Minds: Evaluating Natural and Artificial Intelligence — José Hernández-Orallo, 2017

2. Machine Learning

Enabling machines to learn for themselves. Learning to make predictions and identify patterns given different kinds of data (supervised, unsupervised, and reinforcement learning). Demystifying the objective function.

80/20

- Machine Learning for Humans, Parts 1–3 — Vishal Maini & Samer Sabri, 2017

Deep dive

- Introduction to Statistical Learning, Chapters 1–3 — Gareth James et al., 2013
- AI for Everyone (nine-hour Coursera course) — Andrew Ng, 2019
- The Best Machine Learning Resources — Vishal Maini & Samer Sabri, 2017

3. Deep Learning

Learning to predict an output given an input; drawing inspiration from the brain. Looking inside the black box of deep neural networks. What is a cat, and why? How do we represent the world around us numerically (as sensory "input" or data) and use math to make sense of it? Common architectures.

80/20

- Neural Networks and Deep Learning, Chapter 1 — Michael Nielsen, 2015
- Machine Learning for Humans, Part 4: Neural Networks & Deep Learning — Vishal Maini & Samer Sabri, 2017
- Neural Networks Demystified — YouTube series
- But what *is* a Neural Network? | Deep learning, chapter 1 (video) –3Blue1Brown, 2017
- Architectures — CNN, RNN, LSTM, Transformer

Deep dive

- Neural networks and deep learning — Michael Nielsen, 2015
- Deep Learning Book — Ian Goodfellow, Yoshua Bengio, and Aaron Courville, 2016
- TensorFlow Neural Network Playground

- The building blocks of interpretability — Chris Olah, 2018
- How do neural and symbolic technologies mesh? — Eric Drexler, 2018
- Deep Learning Papers Reading Roadmap — Flood Sung, 2018

4. Deep Reinforcement Learning

Artificial agents learning from reward. Value functions. Exploration and exploitation. Reaching superhuman performance in complex games. Key breakthroughs.

80/20

- Reinforcement Learning: An Introduction — Chapter 1 — Rich Sutton & Andrew Barto, 2017
- Machine Learning for Humans, Part 5: Reinforcement Learning — Vishal Maini & Samer Sabri, 2017
- An intro to Reinforcement Learning (with otters) - Monica Dinculescu, 2018
- NIPS 2017 Keynote: Deep RL Symposium — AlphaZero (video) — Dave Silver, 2017
- AlphaGo Zero — How and Why it Works — Tim Wheeler, 2017

Deep dive

- Reinforcement Learning: An Introduction
- Deep Q-network (DQN) — Mnih et al., 2015
- Deep Reinforcement Learning: Pong from Pixels — Andrej Karpathy, 2016
- AlphaGo, AlphaGo Zero, AlphaZero, AlphaStar — DeepMind
- Spinning Up in Deep RL — OpenAI, 2018
- The Long-Term of AI and Temporal Difference Learning — Richard Sutton, 2017

5. Safety & Alignment

How do we ensure that highly capable and general AI systems reliably understand what we want and help us get it? Identifying and pre-empting potential failure modes. An introduction to specification, robustness, and assurance.

80/20

- Building safe artificial intelligence: specification, robustness, and assurance — DeepMind safety team, 2018
- Superintelligence: Paths, Dangers, Strategies , Ch. 7–10 & 12–13. — Nick Bostrom, 2014
- Benefits & Risks of Artificial Intelligence — Future of Life Institute, 2016
- Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity — Open Philanthropy, 2015
- Faulty Reward Functions in the Wild — OpenAI, 2016

Deep dive

- Takeoff speeds — Paul Christiano, 2018
- Concrete problems in AI safety — OpenAI, 2016
- Scalable agent alignment via reward modeling — Leike et al., 2018
- Reframing Superintelligence: Comprehensive AI Services as General Intelligence — Eric Drexler, 2019
- Fairness, privacy, security & verification resources — Andrew Trask, 2019
- Towards Robust and Verified AI — DeepMind, 2019
- An Overview of Technical AI Alignment (podcast) — Lucas Perry & Rohin Shah, 2019
- AI safety resources — Victoria Krakovna, 2018

6. Strategy & Governance

Near- and long-term strategic challenges and opportunities presented by AI. An introduction to the AI governance problem: the problem of devising global norms, policies, and institutions to best ensure the beneficial development and use of advanced AI.

80/20

- AI governance research agenda — Allan Dafoe, 2018
- 80,000 Hours: Guide to working in AI policy and strategy — Miles Brundage, 2017

- The Malicious Use of Artificial Intelligence — Forecasting, Prevention, and Mitigation
- Asilomar AI Principles — Future of Life Institute, 2017
- Positively shaping the development of artificial intelligence — Robert Wiblin, 2017
- Move 37, or how AI can change the world — George Zarkadakis, 2016

Deep dive

- Bridging near- and long-term concerns about AI - Stephen Cave & Seán S. ÓhÉigeartaigh, 2019
- Deciphering China's AI Dream — Jeff Ding, 2018
- Communication on Artificial Intelligence — European Commission 2018
- Preparing for the Future of Artificial Intelligence — The White House, National Science and Technology Council, Committee on Technology, 2016
- AI and the Economy — NBER Working Paper, Jason Furman & Robert Seamans, 2018
- AI Aftermath Scenarios — Future of Life Institute
- Superintelligence: Paths, Dangers, Strategies, Ch. 14–15 — Nick Bostrom, 2014
- The Vulnerable World Hypothesis — Nick Bostrom, 2018
- Strategic Implications of Openness in AI Development — Nick Bostrom, 2017
- Information Hazards: A Typology of Potential Harms from Knowledge — Nick Bostrom, 2011
- Astronomical Waste: The Opportunity Cost of Delayed Technological Development — Nick Bostrom, 2003
- Thinking About Risks From AI: Accidents, Misuse and Structure — Zwetsloot, Remco and Dafoe (2019)

You can also keep up to date with the latest developments in the AI space by signing up for Import AI by Jack Clark and the Alignment Newsletter by Rohin Shah, and reading them closely every week.

If you enjoyed these resources and are interested in working on the challenges and opportunities presented by artificial intelligence research, check out the 80,000 Hours job board to see who's hiring. If you have questions or feedback, feel free to get in touch.

Thanks to Teddy Collins, Holden Karnofsky, Luke Muehlhauser, Jack Clark, Miles Brundage, Rohin Shah, Emily Oehlsen, Andrew Trask, Jan Leike, Samer Sabri, Amanda Ngo, Jade Leung, and Aleš Flidr for contributing resources and providing feedback on earlier versions of this list.

Artificial Intelligence Machine Learning Ai Safety Ai Governance Ai Strategy