Recommended Resources

This document lists resources that we recommend for people interested in learning more about AI alignment. For more information see, theturingprize.com or contact us at turingprize@gmail.com.

Introductory Resources- *Start here if you're new*

What is AI alignment?

- Four Background Claims (10 mins)

- AGI Safety from First Principles (2 hours)

- AI Racing Toward the Brink (2 hours; podcast)

When should we expect superintelligent AI systems, and why do we think AI alignment is so important?

- [Biological anchors: A trick that might or might not work](Biological anchors) (30 mins)

- Intelligence Explosion: Evidence and Import (30 mins)

- Why AI alignment could be hard with modern machine learning (20 mins)

{ } The Turing Prize

Advanced resources

What do leaders in the field of AI alignment currently believe?
What are the major points of agreement and disagreement?

- [AGI Ruin: A list of lethalities](#) (45 mins)

- [Where I agree and disagree with Eliezer](#) (30 mins)

- [A central AI alignment problem: capabilities generalization, and the sharp left turn](#) (15 mins)

- [How various plans miss the hard bits of the alignment problem](#) (45 mins)

Research resources
How do I get involved in AI alignment research?

- [Alignment research field guide](#) (25 mins)

- [Principles for alignment/agency projects](#) (5 mins)

- [How to get into independent research on alignment/agency](#) (20 mins)

- **If you are interested in AI alignment research,** contact us at turingprize@gmail.com

{ } The Turing Prize