Technical AI Safety Reading List

(adapted from Cambridge Effective Altruism's AGI Safety Fundamentals Syllabus--all credit to them)

Week 1: AGI and superintelligence

What do we mean by artificial general intelligence, and how might we achieve it?

Required readings:

- Three impacts of machine intelligence
- AGI safety from first principles: Superintelligence (first half, ending at the start of the *Paths to Superintelligence* section)
- The Bitter Lesson

Optional readings:

- Shah's summary of Drexler's Reframing Superintelligence report
- Prosaic AI alignment
- AI and compute
- AI and efficiency
- Understanding human intelligence through human limitations
- The power of intelligence

Optional exercises:

- A crucial feature of AGI is that it will possess cognitive skills which are useful across a range of tasks, rather than just the tasks it was trained to perform. Which cognitive skills do humans possess that are useful both in our modern environment, and in the ancestral environment in which we evolved?
- Optional: what are the most likely ways that the hypothesis that we will build AGIs which have transformative impacts on the world might be false?

Week 2: Goals and misalignment

This week we'll focus on how and why AGIs might develop goals that are *misaligned* with those of humans, in particular when they've been trained using machine learning.

Required readings:

- The superintelligent will

- Introduction to Risks from Learned Optimisation
- Will humans build goal-directed agents?

Optional readings:

- Specification gaming: the flip side of AI ingenuity
- Risks from Learned Optimisation: The Inner Alignment Problem
- Gwern on tool AI
- AGI safety from first principles: Goals and agency (first half, ending at the start of *The likelihood of developing highly agentic AGI* section)
- Coherence arguments do not imply goal-directed behaviour
- Risks from Learned Optimisation: Conditions for mesa-optimisation
- Evan Hubinger - podcast on Risks from Learned Optimisation

Optional exercises:

- Hubinger et al. distinguish between *outer* and *inner* alignment problems. Do the examples of specification gaming from the first post qualify as the former or the latter (or both)? What would need to change for you to answer differently?

Week 3: Threat models (Part 1)

Given previous arguments, what might it look like when major problems arise, and how could we prevent them?

Required readings:

- Risks from Learned Optimisation: Deceptive alignment
- What failure looks like

Optional readings:

- Shah's summary of Cotra's report on transformative AI timelines
- Clarifying What failure looks like (part 1)

Optional exercises:

- Christiano's "influence-seeking systems" threat model in *What Failure Looks Like* is in some ways analogous to profit-seeking companies. What are the most important mechanisms preventing companies from catastrophic misbehaviour? Which of those would and wouldn't apply to influence-seeking AIs?
- Optional: what are the individual tasks involved in machine learning research (or some other type of research important for technological progress)? Identify the parts of the process which have already been automated, the parts of the process which seem like they could plausibly soon be automated, and the parts of the process which seem hardest to automate.

Week 4: Threat models (Part 2)

Given previous arguments, what might it look like when major problems arise, and how could we prevent them?

Required readings:

- What multipolar failure looks like
- Optimisation and the intelligence explosion

Optional readings:

- AI alignment landscape
- Takeoff speeds
- Shah's summary of Cotra's report on transformative AI timelines
- Clarifying What failure looks like (part 1)

Optional exercises:

- Christiano's "influence-seeking systems" threat model in *What Failure Looks Like* is in some ways analogous to profit-seeking companies. What are the most important mechanisms preventing companies from catastrophic misbehaviour? Which of those would and wouldn't apply to influence-seeking AIs?
- Optional: what are the individual tasks involved in machine learning research (or some other type of research important for technological progress)? Identify the parts of the process which have already been automated, the parts of the process which seem like they could plausibly soon be automated, and the parts of the process which seem hardest to automate.

Week 5: Learning from humans

This week, we look at three techniques for training AIs based on human data (all listed under "learn from teacher" in Christiano's AI alignment landscape from last week). These are the core building blocks from which techniques for solving outer alignment problems are constructed.

Required readings:

- Imitation learning lecture (part 1 and part 3)
- Deep RL from human preferences (now known as *reward modeling*)
- Flint's summary of Assistance Games

Optional readings:

- Ambitious vs narrow value learning

- Benefits of assistance over reward learning
- Cooperative inverse reinforcement learning
- Reward-rational (implicit) choice: a unifying formalism for reward learning
- Occam's razor is insufficient to infer the preferences of irrational agents
- Alignment Newsletter summary of Human Compatible

Optional exercises:

- Imagine using reward modelling, as described in the second reading from this week, to train an AI to perform a complex task like building a castle in Minecraft. What sort of problems would you encounter?

Week 6: Decomposing tasks for outer alignment

The most prominent research directions in technical AGI safety involve scaling up human-in-the-loop methods by breaking down the process of supervision into subtasks whose correctness we can be confident about. We'll cover three closely-related variants this week (all classed under "build a better teacher" in Christiano's AI alignment landscape).

Required readings:

- An overview of 11 proposals for building safe advanced AI (introduction then proposals 3, 8, and 9. Proposals 2 and 7 may be useful as background for these.)
- Humans consulting HCH
- AI safety via debate blog post

Optional readings:

- Scalable agent alignment via reward modelling (only section 3.2: recursive reward modelling)
- Ajeya Cotra's summary of Iterated Distillation and Amplification
- Supervising strong learners by amplifying weak experts

Optional exercises:

- A complex task like running a factory can be broken down into subtasks in a fairly straightforward way, allowing a large team of workers to perform much better than even an exceptionally talented individual. Describe a task where teams have much less of an advantage over the best individuals. Why doesn't your task benefit as much from being broken down into subtasks?

Week 7: Other paradigms for safety work

A lot of safety work focuses on "shifting the paradigm" of AI research. This week we'll cover three ways in which safety researchers have attempted to do so.

Required readings:

- Embedded agents
- Chris Olah's views on AGI safety
- Open questions in creating safe open-ended AI (first half, ending at the heading *Research Directions for Safe Open-Ended AI*)

Optional readings:

- Open problems in cooperative AI
- Zoom In: an introduction to circuits
- Multi-agent safety
- Emergent tool use from multi-agent interaction
- The rocket alignment problem

Optional exercises:

- Interpretability work on artificial neural networks is closely related to interpretability work on biological neural networks (aka brains). Describe two ways in which the former is easier than the latter, and two ways in which it's harder.
- Optional (for those who are familiar with the POMDP framework): what are the most important disanalogies between POMDPs and the real world?

Week 8: AGI safety in context

In the last week of curriculum content, we'll look at the field of AI governance, as well as more general work on preparing for a future in which artificial minds play a major role.

Required readings:

- Risks from AI: structure, accident, misuse
- GovAI agenda (AI politics section, pages 34-47)

Optional readings:

- The vulnerable world hypothesis
- The windfall clause: distributing the benefits of AI for the common good
- Sharing the world with digital minds
- Cooperation, conflict and transformative AI: preface and sections 1 and 2

Optional exercises:

- In what ways has humanity's response to other threats (e.g. nuclear weapons, pandemics) been better than we would have expected beforehand? In what ways has it been worse? Why?