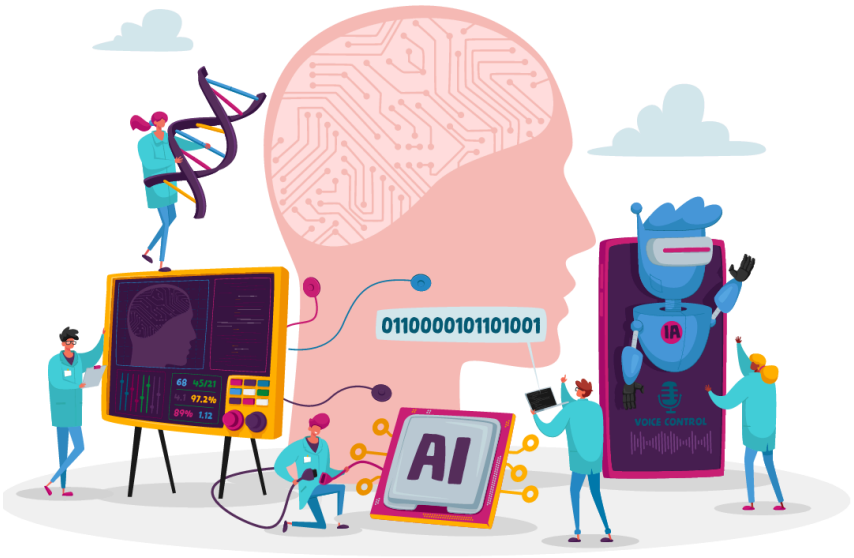
Textbook: AI Safety Fundamentals Express



This document contains all the useful resources for the AISF (previously named AGISF). It will be updated regularly.

This textbook summarizes some key notions and papers in modern AI Safety, requiring minimal familiarity with the other materials covered. We tried to connect the various articles within a unified framework and coherent narrative.

The alignment problem is a pre-paradigmatic problem, and there is no consensus on some of the concepts presented. So please read critically and feel free to add feedback on this form on the various chapters if something seems unclear. Some of the weeks start with a gentle introduction, which you should follow if you are unfamiliar with the concepts.

This textbook is an adaptation of the course held at ENS Paris-Saclay in Paris during March 2023, and offered by EffiSciences and was used and appreciated during the ML4Good bootcamps in Paris and Germany.

Contact

This textbook was created by Charbel-Raphael Segerie, Markov and Jeanne Salle.
Contacts: Charbel-Raphael Segerie , agisf_textbook.0fh6l@slmail.me , jeanne.salle@yahoo.fr

Curriculum

The textbook is currently written as a list of chapters. Some chapters contain a list of quizzes on key resources. Reading carefully each chapter of the textbook should take 1-2 hours.

Week 0: **Artificial Intelligence**

The rest of this textbook assumes familiarity with Machine Learning (ML). View the videos presented in the resources of Week 0 in BlueDot Impact's Alignment course. Please take your time to understand the core concepts, *especially* Reinforcement Learning (RL).
💡Concepts covered: basics of Machine Learning. Basics of Reinforcement Learning, Basics of the Transformer Architecture.

Week 1: **Artificial General Intelligence**

👉Gentle introduction: Artificial General Intelligence (Read the *Characteristics* section)
📖Textbook: Chapter 1.
💡Concepts covered: Current Capabilities, Foundation Models, Leveraging Computation, Future Capabilities, Timelines and Anchors, Instrumental Convergence.

Week 2: **Reward Misspecification**

👉Gentle introduction: An overview of AI Risks by EffiSciences.
📖Textbook: Chapter 2.
💡Concepts covered: Reward, Reward Misspecification, Optimization, Goodhart's Law, Learning by imitation, Learning by Feedback.

Week 3: **Goal Misgeneralization**

👉Gentle introduction: Youtube Video and AI Risks that Could Lead to Catastrophe | CAIS.
📖Textbook: Chapter 3.
💡Concepts covered: Optimization, Mesa-optimization, Inner-Alignment, Deceptive Alignment.

Week 4: **Task Decomposition for Scalable Oversight**

👉Gentle introduction: Youtube Video.
📖Textbook: Chapter 4.
💡Concepts covered: Scalable Oversight Problem, Sandwiching, Task decomposition, Factored cognition, Iterated Amplification, amplification in modern LLMs, Process Supervision.

Week 5: **Adversarial Techniques for Scalable Oversight**

👉Gentle introduction: Youtube Video.
📖Textbook: Chapter 5.1 , Chapter 5.2.
💡Concepts covered:

- Debate, obfuscated arguments problem, AI-written critiques, problems with debate.
- Unrestricted adversarial training, adversarial training, red-teaming language models, interpretability for finding adversarial examples, relaxed adversarial training.

Week 6: **Choose between Agent Foundations and Interpretability**

In week 6 you have a choice between agent foundations and interpretability. If you aren't familiar with interpretability, the vision Interpretability chapter is sufficient. But if you are already familiar with interpretability, you can explore the NLP (natural language processing) interpretability section.

Week 6A: **Interpretability (Default)**

Vision
👉Gentle introduction: Youtube Video.
📖Textbook: Chapter 6.
💡Concepts covered: Feature visualization, saliency techniques (Grad-CAM), activation atlas, circuits, early vision, RL vision, automatic interpretability (NetDissect), multimodal neurons.

Transformers (Bonus, after vision)
📖Textbook: Chapter 6.b

Week 6B: **Agent Foundations**

👉Gentle introduction: Youtube Video.
📖 Textbook: Chapter 6B.
💡Concepts covered: Utility functions, agent AI, tool AI, True Name, Shard theory, CEV, CIRL…

Week 7: **Governance**

📖Please follow Bluedot Impact's course which is already great.

Other resources:

- Detailed list of articles: Alignment Course.
- The old long list of individual summaries: AGISF 2023 Summaries.
- For each week, there is a list of videos for each chapter: Textbook: Videos.
- List of bonus exercises: Archived: 2022 Alignment Fundamentals Curriculum.

The field of alignment does not stop at just the concepts presented in the textbook. Many schools of thought are only briefly touched upon. To get an overview of the field, the best advice is to stay curious and try out various different viewpoints.

One of the best ways to use this textbook is in a reading group consisting of students or young researchers who gather periodically. You don't need to have a large group; a friend is enough, which allows you to discuss the concepts presented in the different chapters critically. If you want to create a reading group to facilitate this course, please get in touch.