# Oxford-Warwick Introduction to AI Safety Workshops – Spring 2021

# Introduction

This syllabus was organised based on that of the 'Safety and Control for Artificial General Intelligence' course at Berkeley, that of Richard Ngo, and that of the Oxford AI Safety Reading Group 2019–2020. Please send feedback and suggestions to Lewis Hammond.

## Signing Up

You can sign up to attend the workshops here (**deadline: 12pm GMT on Sunday 14th February**). They are aimed at those with a somewhat technical background, but not necessarily expertise in AI or AI safety. Applicants should at minimum be seriously interested in learning about AI Safety and able to commit to weekly discussion sessions and readings, and ideally will have already thought about AI safety already and may even be interested in pursuing a relevant career. If this isn't (yet!) you, a good place to get started is this short TED talk by Stuart Russell.

## Practical Details

Sessions will take place at the following times (GMT), and attendees will be sent invites to a video call in advance of each workshop:

- Thursday 18th February 6pm-7.30pm
- Thursday 25th February 6pm-7.30pm
- Thursday 4th March 6pm-7.30pm
- Thursday 11th March 6pm-7.30pm
- Thursday 18th March 6pm-7.30pm

If you have any questions about the practical organisation of workshops, don't hesitate to reach out to Charlotte Siegmann or Rafał Szlendak.

## Preparation

**Before each workshop, please read all of that week's core readings (in bold) and think about the questions** (attempting to write down answers is encouraged but not required). **It is also strongly recommended to read the other readings that are not marked as optional** as these will likely be discussed during sessions. Reading/viewing times for each resource are approximate, and those for papers have been marked with a '+' to indicate that they will likely take more time to read than indicated (unless you are already familiar with the material). **Optional material (either background or further readings) will not be discussed.**

Finally, **don't worry if you don't fully understand things as you're reading them**, the point of the workshops will be to ask questions and resolve uncertainties, as well as discussing the overall themes, features, and implications of the subjects in the material. **You are encouraged to write down questions or things you don't understand and bring them to the workshops.**

# Workshop 1: Introduction to AI and Paths to AGI

## Objectives

- Be aware of AI's historical development and current capabilities/paradigms
- Understand the concept of AGI and its potential impacts
- Be aware of the uncertainty regarding when we will reach AGI and what it will look like
- Understand the concepts of goals and agency as applied to A(G)I

## Optional Background Material

- Artificial Intelligence Wikipedia page (Introduction, Basics, Challenges, and Approaches) – 30 mins
- A Brief Introduction to Reinforcement Learning – 10 mins
- Artificial Intelligence Framework: A Visual Introduction to Machine Learning and AI – 10 mins
- What is AGI? – 5 mins
- Artificial Intelligence: Overview – 20 mins
- 3Blue1Brown's videos on neural networks and gradient descent (both 20 mins) provide excellent intuition and don't require a technical background
- Google's machine learning glossary is also a handy reference

## Progress in AI

- **The Bitter Lesson** – 5 mins
- **When Will AI Exceed Human Performance? Evidence from AI Experts** – 10+ mins
- AI and Compute (excluding appendices) and AI and Efficiency – 10 mins

## AGI

- **AGI safety from first principles: Introduction** – 5 mins
- **AGI safety from first principles: Superintelligence** – 10 mins
- **AGI safety from first principles: Goals and Agency** – 20 mins

## Optional Further Reading

- Three Impacts of Machine Intelligence – 15 mins
- DALL·E (25 mins, but skimmable) is an example of the current state of the art

## Questions

- What do you think the first AGI might look like? Why?
- How does the expected arrival date of AGI impact what work we should do now?
- How likely is it that AGI will be 'goal-directed'?

# Workshop 2: The Alignment Problem and Current Research

## Objectives

- Understand what it means for an AI system to be ‘misaligned’
- Be aware of the possible consequences of powerful misaligned AI systems
- Understand some of the ways in which misalignment might occur
- Gain an overview of current work in AI alignment, and how different problems, proposals, and agendas relate to one another

## The Alignment Problem

- **Specification Gaming: The Flip Side of AI Ingenuity** – 5 mins
- **Clarifying “AI Alignment”** – 5 mins
- **What Failure Looks Like** – 10 mins
- **Risks from Learned Optimization: Introduction** – 15 mins

## Overview

- **Current Work in AI Alignment** (excluding Q&A) – 25 mins

## Optional Further Reading

- Clarifying Some Key Hypotheses in AI Alignment – 10 mins
- AGI safety from first principles: Alignment – 15 mins
- AGI safety from first principles: Control – 15 mins
- AGI safety from first principles: Conclusion – 5 mins
- Concrete Problems in AI Safety – 60+ mins (but only 10+ mins for the Introduction and Overview sections)
- AGI Safety Literature Review – 40+ mins
- An Overview of 11 Proposals for Building Safe Advanced AI – 55 mins

## Questions

- Which problems might we expect to see improved (or conversely, worsened) by more capable AI systems?
- Which current research strands and problems within AI alignment seem most important? Why?
- How tractable is working on alignment now? Can we expect to make progress before we have more powerful AI systems?

# Workshop 3: Value-Learning and Human-Compatibility

## Objectives

- Gain an overview of one of the primary research strands in AI alignment
- Understand assistance games (CIRL) and how they differ from a traditional reinforcement learning setting
- Understand Stuart Russell's 'three principles' for creating safer AI systems

## Overview

- **Summary of Human Compatible: Artificial Intelligence and the Problem of Control** – 10 mins
- **Our Take on CHAI's Research Agenda in Under 1500 Words** – 5 mins
- **Ambitious vs. Narrow Value Learning** – 5 mins

## Cooperative Inverse Reinforcement Learning

- **Cooperatively Learning Human Values** – 10 mins
- Cooperative Inverse Reinforcement Learning – 30+ mins

## Optional Further Reading

- Benefits of Assistance over Reward Learning – 30+ mins
- Inverse Reward Design – 25+ mins
- The Off-Switch Game – 30+ mins
- Human Compatible – 400 mins

## Questions

- What difficulties might we face in learning about human preferences from their behaviour? How could we try and get around these?
- How might assistance games differ when we have multiple humans and multiple AI agents?
- Is narrow value learning sufficient for alignment?

# Workshop 4: Scalable Oversight and Recursive Approaches

## Objectives

- Gain an overview of one of the primary research strands in AI alignment
- Understand the issue of scalable oversight
- Understand the concept of iterated amplification, as well as some proposed instantiations of the overall framework

## Overview

- **Learning from Human Preferences** – 5 mins
- Humans Consulting HCH and Strong HCH – 10 mins
- **Iterated Distillation and Amplification (Summary)** – 10mins

## Recursive Approaches to Alignment

- **AI Safety via Debate** – 10 mins
- **Scalable Agent Alignment via Reward Modeling** – 5 mins
- Supervising Strong Learners by Amplifying Weak Experts – 20+ mins

## Optional Further Reading

- Scalable Agent Alignment via Reward Modeling: A Research Direction – 60+ mins
- AI Safety via Debate – 60+ mins
- Deep Reinforcement Learning from Human Preferences – 25+ mins
- Writeup: Progress on AI Safety via Debate – 50 mins

## Questions

- What are the fundamental assumptions underlying iterated amplification? How likely do you think they are to be true?
- What are the differences (and similarities) between different approaches to iterated amplification, such as debate and reward modelling?

# Workshop 5: Agent Foundations and Decision Theory

## Objectives

- Gain an overview of one of the primary research strands in AI alignment
- Understand the concept of embedded agency, and how it differs from the current paradigms in AI
- Understand the concept of logical uncertainty (in particular, logical counterfactuals and logical priors)
- Understand the concept of corrigibility

## Overview

- **Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda** – 45+ mins

## Topics in Agent Foundations

- Embedded Agency – 10 mins
- Logical Induction (Summary) – 5 mins

## Optional Further Reading

- Corrigibility – 35+ mins
- Toward Idealised Decision Theory – 50+ mins
- Logical Induction (Abridged) – 50+ mins

## Questions

- Which of the 'foundational' problems in MIRI's agenda seem most (or least) important? Why?
- Would it be possible to build aligned AGI without solving these problems? How much more likely would that outcome be if we had solved the problems?
- To what extent do you think practical (i.e., non-theoretical) work on these problems is feasible today?