CUEA Project Based AI Safety Reading Group Syllabus

"It's tempting to dismiss the notion of highly intelligent machines as mere science fiction, but this would be a mistake, and potentially our worst mistake ever."
- Stephen Hawking

"It seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers… At some stage therefore we should have to expect the machines to take control…"
- Alan Turing

"Perhaps the most important thing we can do is to design AI systems that are, to the extent possible, provably safe and beneficial for humans."
- Stuart Russell, Human Compatible

"The more of the control problem that we solve in advance, the better the odds that the transition to the machine intelligence era will go well…. I can imagine that… people a million years from now look back at this century and… say that the one thing we did that really mattered was to get this thing right."
- Nick Bostrom

(If you like quotes, there are more good ones here.)

The field of Artificial Intelligence is progressing rapidly, and advances in AI capabilities raise pressing ethical questions. Many people, including AI experts like Stuart Russell, believe that there is an **urgent need for work** to be done to mitigate the risks associated with future advanced AI systems and to ensure that their contributions are beneficial to humanity and the world.

This discussion group will start by exploring arguments for and against the importance of AI Safety work, especially as it relates to reducing existential risk. The rest of the time, participants will learn about existing AI Safety technical research, efforts to implement policy measures that reduce AI risk, ways they can personally contribute to AI Safety, and more. Many of the readings are quite technical in nature, but if you are willing to try to understand it, you should find most readings accessible without a thorough background in multivariable calculus, linear algebra, and probability (although these will definitely help). The time commitment is about 3-3.5 hours per week: 1 hour of group discussion, and 2-2.5 hours of independent reading/project work. Take your time with the readings and try to really understand them. Let the facilitators know if you have any questions! We are all here to learn together.

The 4 week projects should take around 10-15 hours and will offer you the opportunity to dig into whatever part of the course you found most interesting and exciting! The goal of this project is to help you practice taking an intellectually productive stance towards AGI alignment or governance - to go beyond just reading and discussing existing ideas, and take a tangible step towards contributing to the field yourself. This is a particularly valuable exercise at this point in your exploration of the field because it's so new, still with lots of room to explore.

Note that the fellowship has a **2-week drop policy**. After the second week, if you don't think you'll be able to make the commitment for the next six weeks or don't think the fellowship is for you, you may leave – and this is no problem! We offer this so you can have a graceful exit and remaining fellows get a good discussion environment.

This reading group is a part of Columbia University Effective Altruism (CUEA). For more resources ***please see this extensive reference doc and*** check out CUEA's website!

Week 1 - A Short What, Why, and When of AGI

In the first week, we will start with a basic background of the current state of the art in Artificial Intelligence (AI) and then move on to Artificial General Intelligence (AGI), why it will be important in the future, and when we can expect it to arrive.

We will also be doing an ML review session during the first week for people who aren't as familiar!

Required Readings:

1. Before coming to the meeting:
    a. **What is AI? (30 min) (Skip if you know content already)**
    If you are very unfamiliar with AI and neural networks, please look through a few of these resources. It is crucial you understand the material in the 3Blue1Brown video about neural networks, and would also be beneficial if you understand the rest. We will also have an ML review session to get everyone up to speed.
    b. **The current state of the art in AI (30 min) (Skip if you know content already)**
    If you are not already familiar with the incredible capabilities of modern AI systems (like GPT-3), you should spend about 30 mins exploring the videos and articles that you find at this link.
    c. **What is AGI? (10 min)**
    A short description of what we mean when we say "AGI".
    d. **Clarifying "AI alignment" (10 mins)**
    This post clarifies what Paul Christiano means when he talks about AI alignment, and specifically introduces "intent alignment".
    e. **Forecasting AGI, the "biological anchors" approach summary (30 min)**
    This post presents one approach to predicting when we will have AGI. Timelines for AGI are important for determining what technical and governance related work might be most pressing.
2. **Four background claims (20 min)**
This post has four claims that describe why it might be important to work on AGI safety research.
3. **Superintelligence, Chapter 7: The superintelligent will (35 min)**
Describes some potential issues we will face with super intelligence and introduces common terminology, namely the orthogonality thesis, instrumental convergence, and paperclip maximizers.

Optional Readings:

1. **Introduction to Longtermism (10 min)**
   A short introduction to longtermism, the idea that the influence of decisions and actions of the long-term future is an important moral consideration.
2. **Most important century series summary (15 min)**
   A summary of a series that explains why the 21st century might be the most important in all of human history (a key point is AGI).
3. **Nick Bostrom: What Happens When Our Computers Get Smarter Than We Are? TED Talk (16 min, 2x speed available)**
   Bostrom explains the tremendous influence that AI may have on the future of the world, and some of the broad challenges that need to be addressed to ensure this influence is beneficial.
4. **Holden Karnofsky: AI Timelines: Where the Arguments, and the "Experts," Stand (2021) (2700 words)**
   This piece summarizes several strands of research aimed at forecasting transformative AI, arguing that it will likely come within the lifetimes of young people alive today.

Important Vocabulary and Ideas:

- **Artificial General Intelligence (AGI):** a hypothetical AI which “can understand or learn any intellectual task that a human being can”
- **Narrow AI:** AI designed to solve a single task and which does not solve other tasks
- **Task-Based Approach to AI:** “agents that understand how to do well at many tasks because they have been specifically optimized for each task” (Ngo 2020)
- **Generalization-Based Approach to AI:** “agents which can understand new tasks with little or no task-specific training, by generalizing from previous experience” (Ngo 2020)
- **Transformative AI:** “potential future AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution.” Note that this is not necessarily an artificial general intelligence or superintelligence. (See a more detailed definition at Karnofsky 2016.)
- **Superintelligence:** “an intellect that is much smarter than the best human brains in practically every field” (Bostrom, 1998)

Interesting Discussion Prompts/Ideas:

- What is AGI? Can something be more or less generally intelligent?
- What is general intelligence? Is efficient cross-domain optimization an accurate description?

- What do we think about the four background claims, how should we act based on our understanding of these beliefs?

Week 2 - Basic Goals and Outer Alignment

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:

1. Before coming to the meeting:
    a. **AGI Safety from first principles Section 4.1 first 2 paragraphs (5 min)**
    This short reading will more formally define the idea of outer alignment.
    b. **Specification gaming: the flip side of AI ingenuity (15 mins)**
    This reading describes specification gaming, a type of outer alignment failure.
    c. **Clarifying what failure looks like (25 mins)**
    This post outlines a clarification of a plausible doom scenario proposed originally by Paul Christiano. Try to think if you buy this and how outer and inner alignment failures do or don't play into this scenario!
    d. **Paul Christiano: AI Alignment Landscape (31 mins 2x speed available)**
    An overview of the field of AI Alignment!
    e. **Reinforcement Learning Introduction (10 min) (Skip if you know the content already)**
    An introduction to RL that should give you a basic understanding (see the optional readings for more in depth resources).
2. **Richard Ngo: AGI safety from first principles *Sections 1 (Introduction), 3 (Goals and Agency), and 6 (Conclusion) only* (30 min)**
This is a really great reading that details if agentic AGI should be expected and more claims about the importance of AGI safety.
3. **Is power-seeking AI an existential risk? *Sections 2 and 3 only* (30 min)**
This document details if we should even expect AI to be an existential risk and gives interesting points regarding incentives.

Optional Readings:

1. **Ben Garfinkel: How sure are we about this AI stuff? (30 min, 2x speed available)**
2. **Another Outer Alignment Failure Story (30 min)**
3. **More counter-arguments against taking AGI safety seriously (5-300 min)**
4. **Richard Ngo: AGI safety from first principles (~15000 words total)**

Interesting Discussion Prompts/Ideas:

- What is outer alignment?

- What is agency? Does Richard Ngo leave anything important out?
- Do you agree with the order of confidence that Richard Ngo lists in the conclusion of AGI Safety from First Principles.?
- How does the scenario of “What Failure Looks Like” for AI relate to other risks such as climate change and nuclear war?

Week 3 - Inner Alignment and Proposals for Safe AGI

We will look at a possible failure mode for AGI. Additionally, we will look at several specific research directions in technical AGI safety.

Required Readings:

1. Before coming to the meeting:
    a. **Various proposals for safe advanced AI - Quick summaries (~5 min each)**
    Open these links using Google Chrome, as some of them use the link-to-highlighted-text feature which isn't implemented in other browsers yet. Note that these can blur together if you're not careful, since they are all brief and there are many of them. Consider writing a 1-2 sentence summary of each as a note to yourself to distinguish them in your mind.
        - **Factored Cognition -** Ought post
        - **Approval-based amplification** - Alignment Newsletter summary
        - **STEM AI** - Evan Hubinger's overview
        - **Debate** - OpenAI blog post (excerpt)
        - **Reward modeling** - DeepMind Safety post (excerpt)
        - **Multi-agent safety** - Alignment Newsletter summary
        - **Market making** - Alignment Newsletter summary

    b. If you have time, start taking a look at Risks from Learned Optimization (below), because you likely won't be able to read the whole thing during our meeting.
2. **Risks from Learned Optimization in Advanced Machine Learning Systems (60 min - introduction, 2.0, 3.0, 3.1, and 4.0)**
This paper describes inner alignment and it is quite dense. We encourage you to try to read all of it, but that may take over 60 minutes.

Optional Readings:

1. **Takeoff Speeds (35 min)**
In response to Yudkowsky's (2015) argument that there will be a sharp "intelligence explosion", Christiano argues that the rate of progress will instead increase continuously over time. However, there is less distance between these positions than there may seem: Christiano still expects self-improving AI to eventually cause incredibly rapid growth.
2. **Unsolved problems in ML safety (50 min)**
Hendrycks et al. provide an overview of open problems in safety which focuses more on links to mainstream ML.

3. The easy goal inference problem is still hard **(10 min)**
4. Learning to Summarize with Human Feedback **(15 min)**
5. Reinforcement Learning: An Introduction (Textbook**) (40 mins)**
   I recommend you read the introduction to this introductory textbook if you don't have a good intuitive grasp of the basic ideas of RL.
6. Deep Reinforcement Learning from Human Preferences **(20 min)**
   Read the abstract, the introduction (skip related work), and glance at the results. I also encourage you to try to read section 2 of the paper in detail (this will take way more than 20 minutes). It is good to practice reading papers in detail. Also check out this video showing the results!
7. Iterated Amplification **(30 min)**
   This paper is a very readable paper so hopefully you should be able to read the whole thing!

Week 4 - Understanding AI, Basic AI Governance, and An Overview of the AI Alignment Landscape

This week we will look at agendas and approaches aimed at getting better understandings of the AI systems we are working with. We will also think about strategic considerations around politics that will be important for ensuring AGI goes well. Things would be much easier if there were just one united body with lots of time trying to build aligned AGI, but unfortunately that's not the case.

Required Readings:

1. Before coming to the meeting:
   a. **AGI Week 8: Projects — Effective Altruism Cambridge (eacambridge.org) (5 mins)**
      Using this reading as a guide, **write a ~½ page project proposal** in a Google Doc and share it with your facilitator(s) and cu.effective.altruism@gmail.com. Also feel free to reach out to your facilitator(s) to discuss project ideas. The proposal should describe what you would consider a successful project, and a plan for what to accomplish each week (for 4 weeks) to achieve that successful outcome. Try to take into consideration possible time conflicts in advance. Consider using Murphyjitsu on your plans.
   b. **More proposals for safe advanced AI - Quick summaries**
      Open these links using Google Chrome, as some of them use the link-to-highlighted-text feature which isn't implemented in other browsers yet:
      i. **Microscope AI** - Evan Hubinger's overview
      ii. **Imitative generalization (a.k.a. "Learning the prior")** - Alignment Newsletter summary
   c. **MIRI's approach (Soares, 2015) (25 mins)**
      This reading describes the agenda of the Machine Intelligence Research Institute, namely very foundational, mathematical work on intelligence.
   d. **AI Governance: Opportunity and Theory of Impact (Dafoe, 2020) (25 mins)**
   e. **China's New AI Governance Initiatives Shouldn't Be Ignored (Sheehan, 2022) (10 mins)**
2. **Zoom In: an introduction to circuits (Olah et al., 2020) (35 mins)**
   An introduction to an interpretability approach to reverse engineer neural networks.

3. **Neel Nanda:** **My Overview of the AI Alignment Landscape: A Bird's Eye View (20 min)**

Optional Readings:

1. **Thread: Circuits (200 min)**
A series of short articles building on Zoom In, exploring different circuits in the InceptionV1 vision network.
2. **A mathematical framework for transformer circuits (90 mins)**
Elhage et al. build on previous circuits work to analyze transformers, the neural network architecture used by most of today's cutting-edge models. For a deeper dive into the topic, see the associated videos.
3. **Locating and Editing Factual Associations in GPT (120 mins)**
Bau et al. find a way to change individual associations within a neural network, which allows them to replace specific components of an image. For work along similar lines in language models, see here.
4. **Eliciting Latent Knowledge (up to the end of the Ontology Identification section on page 38) (60 mins)**
This reading outlines the research agenda of Paul Christiano's Alignment Research Center. The problem of eliciting latent knowledge can be seen as a long-term goal of interpretability research.
5. **Cooperation, conflict and transformative AI: sections 1 & 2 (Clifton, 2019) (25 mins)**
6. **Our AI governance grantmaking so far (Muehlhauser, 2020) (15 mins)**
7. **Fischer et al: AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment (Just the 6 page summary, including the table. 20 mins)**
8. **Allan Dafoe: AI Strategy, Policy, and Governance (22 mins, 2x speed available)**
9. Particularly relevant this week is the AGI Safety Fundamentals Governance Track curriculum, which you can find here.

Weeks 5 - 7.9 - Project Work

For the next 4 weeks you have the opportunity to dig into whatever part of the course you found most interesting and exciting! The goal of this project is to help you practice taking an intellectually productive stance towards AGI alignment or governance - to go beyond just reading and discussing existing ideas, and take a tangible step towards contributing to the field yourself. This is a particularly valuable exercise at this point in your exploration of the field because it's so new, still with lots of room to explore. Please plan to dedicate at least 10-15 hours to your project.

Here are some more resources that detail some existing proposals for projects (feel free to copy/copy and slightly modify one of these ideas):

- Project page from AGI Safety Fundamentals
- Project ideas from Buck Shlegeris
- Spreadsheet of project proposals

If you want feedback on a project proposal from a professional researcher in AI Safety, please write a ~one page project proposal on a google doc and share it with cu.effective.altruism@gmail.com and your group's facilitators. We will try to give you feedback within a few days.

Throughout the project phase, our rigid meetings will be more flexible, in that we may do tutorials on PyTorch, in depth explanations of algorithms, status updates, or anything else that would be useful for people's projects.

Week 8 - Project Recaps and AI Safety Landscape and Careers

Please come to week 8 with a ***finished project***! We will also do a brief overview of the rest of AI Safety and discuss careers in the field.

Required Readings

1. **Charlie Rogers-Smith: How to pursue a career in technical AI alignment (spend 30 mins on the parts you find most relevant / interesting)**
   A detailed overview of types of alignment work, advice on choosing between them, guidance on how to pursue them, and thoughts on background learning that will be helpful.

Optional Readings:

1. **Rohin Shah: FAQ: Advice for AI Alignment Researchers (30 min)**
   Just one perspective, but I found this valuable. There are lots of links within this FAQ that you may want to explore as well.
2. **Buck Shlegeris: How I think students should orient to AI safety (first 15 minutes, Q&A after that is optional)**
3. **AGI Further Resources and Organizations (20 min)**
   Really spend about 20 minutes looking through the resources on this document and focus on the section about relevant organizations to get an idea of what other people are doing in the space!