Swarthmore Effective Altruism Society: Introductory Fellowship

What the fellowship involves

Weekly Sessions

Each week of the fellowship has a section of core materials and an optional exercise, which fellows complete in advance of attending the weekly meeting.

We think that the core materials take about one hour to get through, and the optional exercise another 30 minutes - 1 hour. This means that you should expect to spend one or two hours per week in preparation.

We expect Fellows to complete this core reading in advance to help get the most out of the Fellowship and to give a better experience to your peers.

Beyond the required reading, there are more materials each week in 'Recommended Reading' and 'More to Explore.' These are all optional and explore the themes of the week in more depth and breadth. That said, previous fellows have reported getting a lot out of these readings—understanding how to most effectively contribute to the common good is challenging and requires lots of understanding.

Further Opportunities

As part of the fellowship, you will gain access to additional opportunities, including:

- Online socials, including joint events with fellows from other university groups in the US.
- Q&A sessions with individuals from the EA community working in areas relevant to your reading.
- 1-on-1 sessions with fellowship facilitators to discuss your plans and receive personalised advice and connections.

How we hope you'll approach the fellowship[1]

Take ideas seriously.
We think the ideas in this program are important, and worth taking seriously. This means that we think we should be asking ourselves questions like:

- "How could I tell if this idea was true?"
- "If it's true, what does that imply I should be doing differently in my life? What else does it imply I might be wrong about?"
- "If it's untrue, what does that imply I should be doing differently? What else does it imply I might be wrong about?"

And, zooming out:

- "Where are my blind spots?"
- "Which important questions should I be thinking about that I'm not?"

Taking ideas seriously means thoughtfully engaging with questions like these. More generally, it means wanting to make our worldviews as full and accurate as possible, so that we can make better decisions about things that we care about.

Keep an open mind. Part of taking ideas seriously means being open to new ideas and opposing views — even if they initially seem strange or objectionable — and critically engaging with them, as described above. If, after engaging with an idea we didn't initially agree with, we find that there's a compelling case for it, we should be open to changing our minds. This is especially important because the issues that we'll discuss in this program are generally very complicated, and it's often really difficult to understand them fully and truthfully (especially when we can't necessarily rely on conventional wisdom). This open-mindedness (and the thoughtful disagreement, re-evaluation, and questioning that comes with it) is important to understanding EA, as a community with lots of disagreement about many of the issues it focuses on.

Disagreements are interesting. When thoughtful people with access to the same information reach very different conclusions from each other, we should be curious about why. If, for example, a medical community is divided on whether Treatment A or B does a better job of curing some disease, they should want to get to the bottom of that disagreement, because the right answer matters — lives are at stake. Even if you don't expect to end up agreeing with the other person, it's valuable to try to *understand* why you disagree, and exactly how their views differ from yours.

[1] Inspired by Julia Galef's Update Project

Try to make guesses and think through things, even if you might be wrong. We'd prefer that you express your feelings, thoughts, and opinions despite being unsure, since saying "I don't know" or making vague claims can be a way to hide important disagreements or avoid thinking through the implications of ideas. In practice, this might sound something like: "I haven't looked into this much, but I've read articles that make me think that people living in hunter-gatherer communities are healthier, and maybe happier, than people living in industrialized societies." It's easy to have a vague disagreement, and just "agree to disagree," without ever learning from each other. A clearly expressed view makes it easier for other people to figure out exactly what you might disagree about, and we think you can mostly usefully learn from others when your disagreements become concrete.

Be aware of our privilege and the sensitiveness of these issues. We shouldn't lose sight of our privilege in being able to discuss these ideas, and shouldn't let that overshadow the fact that we are talking about real lives. We should be aware that we're lucky to be in a position where we can have such a large impact, and that this opportunity for impact is the consequence of a profoundly unequal world. In addition, some of these topics can be uncomfortable to talk about—which is one of the reasons they're so neglected, and so important to talk about—especially when many of us may have personal ties to some of these areas.

Logistics and Communication

You should have received an invitation to join the Fellowship Slack. All communication between fellows and facilitators will take place here.

You will be able to communicate with your cohort facilitator in a private channel created for your cohort. **Please let your facilitator know in advance if you will be unable to attend a session.** If you are unable to attend your session, we recommend that you ask your facilitator about joining another cohort for the session you missed.

.

Week 1: Triage and the Scale of Suffering

"We are always in triage. I fervently hope that one day we will be able to save everyone. In the meantime, it is irresponsible to pretend that we aren't making life and death decisions with the allocation of our resources. Pretending there is no choice only makes our decisions worse."
— Holly Elmore

Over the course of Week 1 and 2 we aim to introduce you to the core principles of Effective Altruism. This week we'll investigate what opportunities to do good we have available to us; come to terms with the tradeoffs we face in our altruistic efforts; and explore tools that can help us find unusually high impact opportunities.

Core Materials

- Global poverty and the demands of morality - Toby Ord (15 mins.)
- Doing Good Better - William MacAskill - Introduction (10 mins.) - *An example of two public policy interventions that have massively different impacts.*
- Introduction to EA | Ajeya Cotra | EAGxBerkeley 2016 - *An excellent introductory talk delivered to students at Berkeley* (30 mins., or 15 mins. at 2x speed)
- We are in triage every second of every day - Holly Elmore (5 mins.)
- 500 Million, But Not a Single One More (5 mins.)
- The world is much better; The world is awful; The world can be much better - Our World in Data (5 mins.)

Recommended reading

- Famine, Affluence, and Morality (p. 229-236) - Peter Singer (15 mins.) - *An introduction to the famous "drowning child" thought experiment in moral philosophy*
- The lack of controversy over well-targeted aid - GiveWell (10 mins.)
- Excited altruism - GiveWell - *Where does our own passion and excitement fit into the picture?* (10 mins.)

More to explore

- Save a life or receive cash? Which do recipients want? - IDinsight - *Explores the preferences and values of individuals and communities in Ghana and Kenya to inform funding allocations.* (10 mins.)
- Q&A with Elie Hassenfeld *Elie, the CEO and co-founder of GiveWell, discusses his organization's latest research, his views on economic growth, and what he's changed his mind on lately* (60 mins.)

- Reasoning Transparency - Open Philanthropy Project - *A guide on how to write while being more transparent on the reasoning behind your views.* (25 mins.)

Week 2: How Big Are Differences in Impact?

"Most decisions of this sort take dramatically insufficient account of cost-effectiveness. As a result, thousands or millions of people die who otherwise would have lived. The few are saved at the expense of the many. It is typically done out of ignorance about the significance of the cost-effectiveness landscape rather than out of prejudice, but the effects are equally serious."
— Toby Ord

In Week 2 we continue to explore the core principles of Effective Altruism. We focus on giving you tools to quantify and evaluate how much good an intervention can achieve; introduce important tools such as expected value reasoning and scope sensitivity; and investigate differences in expected cost-effectiveness between interventions.

Organisation Spotlight

GiveWell

GiveWell searches for the nonprofits that save or improve lives the most per dollar. They recommend a small number of non-profits that they believe do an incredible amount of good. Unlike non-profit evaluators that focus solely on financials, assessing administrative or fundraising costs, they conduct in-depth research aiming to determine how much good a given program accomplishes (in terms of lives saved, lives improved, etc.) per dollar spent.

Rather than try to rate as many non-profits as possible, they focus on the few non-profits that stand out most (by their criteria) in order to find and confidently recommend high-impact giving opportunities. See their top non-profit here.

They believe that there is exceptionally strong evidence for their top non-profit, and that donations can save a life for every $3,000-$5,000 donated.

Core Materials

- Doing Good Better - William MacAskill - Chapters 5, 6, and 7 (45 mins.)
- On Caring (15 mins.)
- The Moral Imperative toward Cost-Effectiveness in Global Health - Centre for Global Development (20 mins.)
- Expected Value (2 mins.)
- Scope Insensitivity (3 mins.)

Recommended Reading

- Global Poverty and the Demands of Morality - Toby Ord (20 mins.)
- What are the most important moral problems of our time? - Will MacAskill - (12 mins.)
- One approach to comparing global problems in terms of expected impact - 80,000 Hours - *An outline of a more precise and quantitative version of the importance, neglectedness, and tractability framework; and details on how to apply it to make your own comparisons of areas.* (20 mins.)
- Our Criteria for Top Charities - GiveWell and Process for Identifying Top Charities - GiveWell (20 mins.)
- Our current list of especially pressing world problems - 80,000 Hours (20 mins.)

More to Explore

- Scope insensitivity: failing to appreciate the numbers of those who need our help - *A psychological explanation of scope insensitivity and its implications for the cause of animal welfare.* (5 mins.)
- Our Use of Cost-Effectiveness Estimates - Animal Charity Evaluators - *A description of the role that cost-effectiveness estimates play in ACE's intervention reports and charity evaluations, as well as the challenges of creating cost-effectiveness estimates, and the risks of making them public.*(40 mins.)
- List of ways in which cost-effectiveness estimates can be misleading - *A checklist of things to keep in mind when using cost-effectiveness estimates.* (25 mins.)
- Prospecting for Gold - *An overview of potential methodological tools we can use to find and evaluate high-impact opportunities.* (Video - 55 mins.)

Optional Exercise (60 mins.)

Your challenge this week will be to attempt to estimate how much impact you might be able to achieve by donating to effective charities.

Part 1 - estimate your likely total future income (10 mins.)

In this part of the exercise we'd like you to estimate what your total future income will be during your life. This is obviously quite a personal question, so this estimate is just for you, and we won't be explicitly discussing answers to this part in-session. If you'd feel more comfortable, feel free to just estimate what an average graduate from your college will earn.

In making this estimate, of course, it's difficult to know what this will be and impossible to know what the future holds for you. But we think that you still might be able to make educated estimates based on factors such as what the typical graduate earns, what you think your likely career paths are that you're considering now, and sense checking the answer.

Feel free to do any research that you would like to make your estimate.

If you're feeling stuck, here are some tips:

- Sometimes life can throw you curveballs and mess up your plans. Try making worst case scenario, most likely scenario, and best case scenario estimates if you're feeling uncertain about what the future holds
- Break the question down eg. you might find it useful to start by estimating how many years you'll work before retirement
- Don't worry about this question too much and try not to spend more than 10 to 15 minutes on it. It's okay to just go with a very rough and inaccurate guess.

Write your answer here

We're trying to put a number on the total income you'll earn over the course of your life

You might plug in this value into Giving What We Can's How Rich Am I? calculator to see how this average annual income compares to the rest of the world.

Part 2 - what could you achieve with your income? (20 mins.)

For the second part of the exercise we'll try and work out what you could achieve by donating some of the money you'll earn in the future.

GiveWell is an effective altruism inspired organisation which attempts to identify outstanding donation opportunities in global health and development. Using their reports on their top charities[2] and your earlier estimate of your future income, try and work out what you could achieve if you donated 10% of your lifetime income to one of these charities.[3] If you'd like to explore further, check out GiveWell's cost effectiveness models.

Complete this exercise for three GiveWell charities.

Write your answer here

[2] If you're short on time, here's a cheat sheet with information about three top GiveWell charities

[3] 10% is the figure of the Giving What We Can Pledge, a pledge that many involved in the Effective Altruism community have taken.

e.g.

Malaria Consortium: X months of malaria prevention for one person, with an estimate of N deaths averted

GiveDirectly: $X transferred to recipients

Part 2b (10 mins.)

In the last section, you ended up with a few different options, (e.g. saving the lives of 200 5-year olds, doubling the income of 1000 people earning $1/day, or a 10% chance of preventing 1000 kids from attending school). Now imagine you get to donate to one of these charities.

There's a difficult judgement to be made now: since you have to pick, which charity would you donate to to do the most good? The Fellowship will donate $100 to whichever charity has the most votes at the end of this week.

Write your answer here

Which charity do you pick to donate to? Why?

Optional (10 mins.)

What are other decisions in your life that you might consider generating quantitative estimates and comparing outcomes for?

(Optional) Jot your thoughts down here

Week 3: Expanding the Moral Circle

"The question is not, Can they reason?, nor Can they talk? but, Can they suffer? Why should the law refuse its protection to any sensitive being?"
— Jeremy Bentham (1789)

This week focuses on a core idea in effective altruism: that one way to make a large positive impact is by helping those who mainstream society neglects, and that finding out where. Historically, it has always been the case that past generations were (sometimes unwittingly) complicit in what we would retrospectively call moral atrocities—the disregard of foreigners, the subjection of women, the enslavement of entire peoples—due in part to the view that these groups were not owed moral concern. During Week 3 we explore who our moral consideration should expand to, with a particular focus on farmed animals as a case example.

Organisation Spotlight

Animal Charity Evaluators

Animal Charity Evaluators (ACE) aims to identify the best ways to help animals as effectively as possible. They strive to identify ways to alleviate suffering and improve the lives of animals on a wide scale, while continuously updating their recommendations based on new evidence.

Based on their current findings, they believe advocating for farmed animals seems to be the most effective way to help animals and prevent the largest amount of suffering.

Their recommended charities use a range of strategies to help animals including corporate outreach, legal work, and developing alternative proteins. Interventions in this area can be surprisingly cost-effective, for example it seems that, even under conservative estimates, tens or hundreds of chickens can be spared from cage confinement per dollar spent (Source). As of January 2020, ACE has influenced over $26 million in donations.

Core Materials

- Radical Empathy - Open Philanthropy Project (10 mins.)
- Moral Progress and Cause X (5 mins.)
- All Animals Are Equal - *The opening chapter of* Animal Liberation *(1975), widely regarded as the founding text of the animal rights movement.* (25 mins.)
- *The Expanding Circle* pg. 111-124 'Expanding the Circle of Ethics' section (20 mins.)

- Dominion - *Dominion uses drones, hidden and handheld cameras to document the reality of factory farming for food production.* (Film - Just try watching it, for as long as you feel able to. We don't expect you to finish the entire film.)
  - **Content Warning:** Much of the film here can be extremely disturbing and includes graphically violent footage of factory farming. Please make sure to watch this in a moment without e.g. any upcoming deadlines or important meetings the same day. We include it because we think it's important to really see how broken the world is.

Recommended reading

- The next global agricultural revolution - *Bruce Friedrich of the Good Food Institute on how alternative proteins like plant-based and cell-based meat can end factory farming (Video 5 mins.)*
- On "fringe" ideas - *What does it take to keep ourselves open to new possibilities in what the most important problems in the world are?* (10 mins.)
- The Case Against Speciesism - Centre for Reducing Suffering (10 mins.)
- Animal Welfare (20 mins.)
- The Possibility of an Ongoing Moral Catastrophe (30 mins.)
- Should animals, plants, and robots have the same rights as you? - Vox (20 mins.)
- Expanding the moral circle - Sentience Institute (25 mins.)
- *Animal Liberation*, Chapter 3 - Down on the factory farm (60 mins.)
- Suffering in Animals vs. Humans (13 mins.)

More to explore

- Our descendants will probably see us as moral monsters. What should we do about that? - 80,000 Hours - *A podcast featuring Professor Will MacAskill about what we should do if we are making major moral mistakes today.* (Podcast - 1h 50m)
- Practical ethics given moral uncertainty - *Will MacAskill presents a framework for making decisions under moral uncertainty, and offers some implications of this* (5 mins.)
- Social Movement Lessons From the British Antislavery Movement - Sentience Institute - *This report aims to assess (1) what factors led the British government to abolish the transatlantic Slave trade in 1807 and then human chattel slavery in 1833, and (2) what those findings suggest about how modern social movements should strategize.* (2.5 hours)
- The Importance of Wild-Animal Suffering - Centre on Long-Term Risk - *An argument for us to take into account the wellbeing of animals that live in the wild.* (40 mins.)
- The Narrowing Circle (see here for summary and discussion) - *An argument that the "expanding circle" historical thesis ignores all instances in which modern ethics narrowed the set of beings to be morally regarded, often backing its exclusion by asserting their non-existence, and thus assumes its conclusion.* (30 mins.)
- 2017 Report on Consciousness and Moral Patienthood - An *investigation into what types of beings merit moral concern.* (6 hours, skimmable)
- The Subjection of Women - *An essay published in 1869 by John Stuart Mill, with ideas he developed with his wife Harriet Taylor Mill arguing for the emancipation of women* (10 mins.)
- Loving-Kindness and Compassion Meditation: Potential for Psychological Interventions - *A widely cited study into meditation based psychological practices to increase kindness and compassion.* (40 mins.)

- Ethical.diet - *A tool that explains which diet changes have the biggest effects on animal welfare.*

Exercise (25 mins.)

This week's exercises are about doing some personal reflection. There are no right or wrong answers here, instead this is an opportunity for you to take some time and think about your ethical values and beliefs.

Part 1 - A letter to the past (10 mins.)

"Imagine effective altruism had existed at a different point in history. Would the movement have been able to do any good, or would it have been too stuck in the assumptions of the time period?

Would an effective altruist movement in the 1840s U.S. have been abolitionist? If you think such a movement would have failed to stand up against slavery, what do we need to change, now, as a movement, to make sure we're not getting similarly big things wrong?

Would an effective altruist movement in the 1920s U.S. have been eugenicist? If you think the movement would have embraced a pseudoscientific and deeply harmful movement like the sterilization campaigns of the Progressive era, what habits of mind and thought would have prevented us from doing that, and are we actively employing them?

Imagine someone walked into that 1840s EA group and said, "I think black people are exactly as valuable as white people and it should be illegal to discriminate against them at all," or someone walked into the 1920s EA group and said, "I think gay rights are really important." I want us to be a community that wouldn't have kicked them out."

- On "fringe" ideas - Kelsey Piper, edited

Imagine someone from the past who lived at a different time and held views characteristic of that time. Also imagine, for the sake of the exercise, that this person is not too different from you - perhaps you would've been friends. Unfortunately, most people in the past were complicit in horrible things, such as slavery, sexism, racism, and homophobia, which were even more prevalent in the past than they are now. And, sadly, this historical counterpart is also complicit in some moral tragedy common to their time, perhaps not out of malevolence or ill-will, but merely through indifference or ignorance.

This exercise is to write a letter to this historical friend arguing that they should expand their moral circle to include a specific group that your present self values. Imagine that they are complicit in owning slaves, or in the oppression of women, people of other races, or sexual minorities.

For the sake of this exercise, imagine your historical counterpart is not malevolent or selfish, they think they are living a normal moral life, but are unaware of where they are going wrong. What could you say to them to make them realise that they're doing wrong? What evidence are they overlooking that allows them to hold their discriminatory views? You might want to write a few paragraphs or just bullet points, and spend time reflecting on what you write.

Write your letter here

Part 2 - A letter from your future self (15 mins.)

Now imagine one day you get a strange letter; it's a letter from your future counterpart, hundreds of years in the future. In the letter they argue that, just like your past counterpart, you currently are unknowingly and unwittingly committing some moral wrong.

What do you think the letter might say? What issue might be of great moral importance that you are unaware of today?

Again, you might want to write a few paragraphs, and spend some time reflecting on what you write.

Write your letter here

Week 4: What We Owe the Future

"If all goes well, human history is just beginning. Humanity is about two hundred thousand years old. But the Earth will remain habitable for hundreds of millions more—enough time for millions of future generations; enough to end disease, poverty and injustice forever; enough to create heights of flourishing unimaginable today. And if we could learn to reach out further into the cosmos, we could have more time yet: trillions of years, to explore billions of worlds. Such a lifespan places present-day humanity in its earliest infancy. A vast and extraordinary adulthood awaits."
— Toby Ord

In Weeks 1 and 2 we discussed attempting to quantify the impact of altruistic interventions. However, most cost-effectiveness analyses can only take into account the short-run effects of the interventions, and struggle to take into account long-run knock-on effects and side effects. This criticism has been made forcefully against early effective altruist attempts to evaluate interventions based on cost-effectiveness.

This week we'll explore a different approach to finding high-impact interventions - 'longtermism' - which attempts to find interventions that beneficially influence the long-run course of humanity.

Organisation Spotlight

All-Party Parliamentary Group for Future Generations

The All-Party Parliamentary Group (APPG) for Future Generations is a UK parliamentary group working to create cross-party dialogue on combating short-termism and identifying ways to internalise concern for future generations into today's policy making.

They believe that political short-termism can cause topics with widespread consequences – like climate change, public health trends and catastrophic and existential risks – to be neglected from the political agenda in favour of urgent matters.

You can see their research aimed at informing Parliamentarians on catastrophic risks and potential policy options here. You can see their events bringing together policy, academic, and industry communities here.

Required Materials

- What We Owe the Future (Video, 40 mins. or 20 mins. at 2x speed.)

- *The Precipice* - Introduction and Chapter 1 (~40 mins.)
- Why I Find Longtermism Hard, and What Keeps Me Motivated - (10 mins.)

Recommended reading

- All Possible Views About Humanity's Future Are Wild - (15 mins.)
- Can We Predictably Improve the Far Future? (Video - 30 mins.)
- Orienting towards the long-term future (Video - 25 mins.)
- The Case for Strong Longtermism - Global Priorities Institute (1hr. 20 mins.)
- Policymaking for Posterity - (60 mins.)

More to explore

- Representing future generations - *Political institutions generally operate on 2-to-4-year timescales which aren't long enough to address global issues (as the issue of climate change has shown). This talk analyzes sources of political short-termism and describes institutional reforms to align government incentives with the interests of all generations.* (Video - 30 mins.)
- Climate Change and Intergenerational Justice - UNICEF - *How should we balance the rights of those alive today with the rights of future generations?* (10 mins.)
- How becoming a 'patient philanthropist' could allow you to do far more good - *A researcher from the Global Priorities Institute explains how investing resources, instead of spending them immediately, can allow us to do much more good.* (Podcast - 2.5hr)
- Blueprints (& lenses) for longtermist decision-making - *How are we supposed to apply longtermism in practice? The author outlines two concepts of a 'blueprint' and a 'lens' to clarify this issue.* (7 mins.)

Week 5: Existential Risk and the Future of Humanity

"So if we drop the baton, succumbing to an existential catastrophe, we would fail our ancestors in a multitude of ways. We would fail to achieve the dreams they hoped for; we would betray the trust they placed in us, their heirs; and we would fail in any duty we had to pay forward the work they did for us. To neglect existential risk might thus be to wrong not only the people of the future, but the people of the past."
—Toby Ord

As a consequence of the force of longtermist arguments explored last week, some in the effective altruism community have argued that safeguarding the future of humanity from extinction should be among our very top priorities as people trying to help others. In this week, we introduce the concept of existential risk and take a look at some of the specific scenarios and technologies that give most worry to those concerned with the future of humanity.

Organisation Spotlight

Future of Humanity Institute

The Future of Humanity Institute (FHI) is a multidisciplinary research institute working on big picture questions for human civilisation and exploring what can be done now to ensure a flourishing long-term future.

Future of Humanity Institute
UNIVERSITY OF OXFORD

Currently, their four main research areas are:

- **Macrostrategy** - investigating which crucial considerations are shaping what is at stake for the future of humanity
- **Governance of AI** - understanding how geopolitics, governance structure, and strategic trends will affect the development of advanced artificial intelligence
- **AI Safety** - researching computer science techniques for building safer artificially intelligent systems
- **Biosecurity** - working with institutions around the world to reduce risks from especially dangerous pathogens

Required Materials

- *The Precipice*, Chapter 2 - Existential Risk (30 mins),
- *The Precipice* - Chapter 5 (pages 121-138) - Pandemics (25 min.)
- The case for taking AI seriously as a threat to humanity (10 min)
- Policy and research ideas to reduce existential risk - 80,000 Hours (5 mins.)

Recommended reading

- On future people looking back at 21st century longtermism (15 mins.)
- *The Precipice*, Chapter 4 - Anthropogenic Risks (65 mins) *On the risks posed by climate change, nuclear war, and environmental degradation.*
- Reducing Global Catastrophic Biological Risks Problem Profile - 80,000 Hours (60 mins.)
- Professor Stuart Russell on the flaws that make today's AI architecture unsafe & a new approach that could fix it (2 hr. podcast episode)

More to explore

- *The Precipice* - Chapter 3 Natural Risks- *How big is the threat to humanity posed by asteroids and comets, supervolcanoes, stellar explosions, and other natural risks?* (60 mins.)
- The Vulnerable World Hypothesis - Future of Humanity Institute - *Scientific and technological progress might change people's capabilities or incentives in ways that would destabilize civilization. This paper introduces the concept of a vulnerable world: roughly, one in which there is some level of technological development at which civilization almost certainly gets devastated by default.* (45 mins.)
- Open until dangerous: the case for reforming research to reduce global catastrophic risk (Video - 50 mins.)
- S-risks (risks of extreme suffering): Why they are the worst existential risks, and how to prevent them (20 mins.) - *An argument for why risks of astronomical suffering may be the most important existential risks.*

Artificial Intelligence

- What is artificial intelligence? Your AI questions, answered - Vox (40 mins.)
- *The Precipice* - Chapter 5 (pages 138-152) - Unaligned Artificial Intelligence (25 min.)
- What Failure Looks Like (12 minutes) - *Two specific stories about what a very bad society-wide AI alignment failure could look like, which differ considerably from the classic "intelligence explosion" story*
- Some Background on Our Views Regarding Advanced Artificial Intelligence - Open Philanthropy Project - *An explication of why there is a serious possibility that progress in artificial intelligence could precipitate a transition comparable to the Neolithic and Industrial revolutions.* (60 mins.)
- *Human Compatible: Artificial Intelligence and The Problem of Control* (Book)
- *The Alignment Problem: Machine Learning and Human Values* (Book)

Pandemics and Biological Risks

- Dr Greg Lewis on COVID-19 & the importance of reducing global catastrophic biological risks
- Global Catastrophic Risks Chapter 20 - Biotechnology and Biosecurity Biotechnological power is increasing exponentially, at a rate as fast or faster than that of Moore's law, as measured by the time needed to synthesise a certain sequence of DNA. This has important implications for biosecurity. (60 mins.)

Global governance and international peace

- Ambassador Bonnie Jenkins on 8 years of combating WMD terrorism - *an interview with Bonnie Jenkins, Ambassador at the U.S. Department of State under the Obama administration, where she worked for eight years as Coordinator for Threat Reduction Programs in the Bureau of International Security and Nonproliferation.* (Podcast - 1 hour 40 mins.)
- Why effective altruists should care about global governance - *Because global catastrophic risks transcend national borders, we need new global solutions that our current systems of global governance struggle to deliver.* (Video - 20 mins.)
- Destined for War: Can America and China Escape Thucydides's Trap (Book)

Climate Change

- Climate Change Problem Profile - 80,000 Hours - *An analysis of the worst risks of climate change, and some of the most promising ways to reduce those risks.* (30 mins.)
- What can a technologist do about climate change? - *A wide collection of technical projects to reduce the burning of fossil fuels.* (60 mins.)

Nuclear security

- Daniel Ellsberg on the creation of nuclear doomsday machines - *Daniel Ellsberg on the institutional insanity that maintains large nuclear arsenals, and a practical plan for dismantling them* (Podcast - 2 hours 45 mins.)
- List of nuclear close calls - Wikipedia - *A description of the thirteen events in human history so far that could have led to an unintended nuclear detonation* (5 mins.)

Other

- Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority - Centre for a New American Security - *An argument for how advances in military technology (including but not limited to AI) can impede relevant decision making and create risk, thus demanding greater attention by the national security establishment.* (60 mins.)
- Big nanotech: towards post-industrial manufacturing - The Guardian - *an explanation of how atomically precise manufacturing could displace industrial production technologies and bring radical improvements in production cost, scope, and resource efficiency.* (10 mins.)
- AlphaGo - The Movie - DeepMind - *A documentary exploring what artificial intelligence can reveal about the 3000-year-old game of Go, and what that can teach us about the future potential of artificial intelligence.* (Video - 1hr. 30 mins.)
- The Artificial Intelligence Revolution: Part 1 - *A fun and interesting exploration of artificial intelligence by the popular blogger Tim Urban.* (45 mins.)
- The Future of Surveillance - *An exploration of ways in which the future of surveillance could be bad, and an investigation into accountable, privacy preserving surveillance protocols.* (Video - 15 mins.)

Week 6: EA's Criticisms and Internal Debates

"MacAskill does not address the deep sources of global misery – international trade and finance, debt, nationalism, imperialism, racial and gender-based subordination, war, environmental degradation, corruption, exploitation of labour – or the forces that ensure its reproduction. Effective altruism doesn't try to understand how power works, except to better align itself with it. In this sense it leaves everything just as it is. This is no doubt comforting to those who enjoy the status quo – and may in part account for the movement's success."
—Amia Srinivasan

This week, we'll read and discuss critiques of effective altruism, and criticisms of how some people try to implement EA. We are dedicating a week to this because, to whatever extent we are wrong, it would be good to know. Honestly reckoning with strong counter arguments (from both within and outside of the EA community) can help us avoid confirmation bias and groupthink, and get us a little closer to identifying the most effective ways to do good. Such critiques have led to important changes in what many EAs do: for example, many EAs now prioritize longtermism because strong arguments against short-termism were made, and GiveWell polled a sample of its recipients on how they would make moral tradeoffs in response to criticisms that it shouldn't make moral tradeoffs on behalf of the people its recommended charities benefit.

Organisation Spotlight

Global Priorities Institute

The Global Priorities Institute (GPI) is a multidisciplinary research institute. It conducts foundational research to inform the decision-making of those seeking to do as much good as possible. The institute seeks to ensure that the ideas of EA and their applications can withstand intellectual criticism.

GPI's research areas include:

- Assessing the idea of longtermism and its applications
- Investigating value questions of how people should act when they are uncertain about what will happen and what is right
- Examining the relationships between economic growth and well-being

Required Materials

- Evidence, Cluelessness, and the Long Term - Hilary Greaves - (30 mins.)
- Just Giving: Why Philanthropy is Failing Democracy and How It Can Do Better (excerpt) - Rob Reich (10 mins.)
- Stop the Robot Apocalypse - Amia Srinivasan - (15 mins.)
- Pascal's Mugging (~5 min) *A critique of the application of expected value theory to disastrous events with very low probability, which may be relevant to existential risk reduction efforts.*

Recommended reading

- A critique of effective altruism - *A thorough criticism of EA written by an effective altruist trying to challenge their own ideas.* (11 mins.)
- Objections to Value-Alignment among Effective Altruists - *Carla Zoe Cremer argues that certain trends toward intellectual homogeneity in EA may stifle long-term potential of the movement.*
- Effective altruists love systemic change - *Robert Wiblin argues why EA does not, in fact, neglect systemic change.* (13 mins.
- Some personal thoughts on systemic change - *A reflection on how EAs think about systemic change; see comments for disagreement and discussion.*
- Another Critique of Effective Altruism - *Written in the same vein as the above, but covering a few points the other post may have missed.* (5 mins.)

More to explore

- Why we can't take expected value estimates literally (even when they're unbiased) - *Holden Karnofsky explains why he takes issue with using expected value estimates of impact.* (35 mins. - skimmable)
- The Repugnant Conclusion - *A conclusion that total utilitarianism leads to (maximizing overall wellbeing over all beings requires that many many beings with infinitesimally positive wellbeing to be preferred to a smaller number of beings that are all extremely well off, which doesn't seem intuitive).* (6 min. video)
- Ethical Systems - *Check out other ethical systems not discussed yet in the program. Which ones resonate most with you?* (Varies)
- How not to be a "white in shining armor" - *How GiveWell (as of 2012) tries to avoid "developed-world savior" interventions that don't take into account local context* (3 min)
- AI alignment, philosophical pluralism, and the relevance of non-Western philosophy - *Short talk with Center for Effective Altruism* (18 min)
- Making decisions under moral uncertainty - *Placing credence in multiple ethical systems leads to questions of moral uncertainty, when the two ethical systems disagree. This post summarises the problem and suggests ways to resolve such issues.* (16 mins.)
- Utility monster - *Another thought experiment suggesting that trying to maximize well-being may have counterintuitive implications*
- Some blindspots in rationality and effective altruism - *An EA forum blog post that discusses some common pitfalls for rationalists and effective altruists, as well as some meta-considerations* (12 min)

Exercise (10 mins.)

Over the last few weeks we've covered a lot of material. Ethical and moral philosophy foundations of effective altruism, ways of thinking and frameworks for comparing between causes and determining the best way to direct our resources and actions, and some top priority causes using the EA framework.

What are your biggest questions, concerns, and criticisms based on what we've discussed so far? These can be about the EA framework/community, specific ideas or causes, anything you'd like!

Please bring them up and discuss them at your next meeting!

Write your answer here (In your own copy of the document)

Week 7: In What Causes Can We Help Others the Most?

"We think the most important single factor determining the expected impact of your work in the long term is probably the issue you choose to focus on. For example, you might choose to focus on climate change, education, technological development, or something else. We think it's of paramount importance to choose carefully."
—Ben Todd and the 80,000 Hours team

There are so many global problems that could use more attention, and they all interact. We can imagine the ideal 'world portfolio' of effort: if society were perfectly optimised for a flourishing civilization, how would people and resources be spread across different global issues?

None of us have control of the world portfolio, so as individuals, the best we can do is to try to find a way to use our strengths to address one of its biggest gaps, and help the world take a small step towards the ideal.

Unfortunately, the world portfolio is nowhere near the optimal allocation: some issues are far bigger than others, some receive many more resources, and these do not perfectly line up.

When we step back and ask which gaps seem like they most need filling right now – it seems that some are over 100 times more pressing than others, and these others include many of the issues people typically work on.

Organisation Spotlight

80,000 Hours

80,000 Hours seeks to identify careers that have an exceptional social impact, publishes research on priority career paths and problem areas, and provides one-on-one career advising to readers looking to make a large positive impact with their careers. Founded in 2012, 80,000 Hours has collaborated with researchers at the University of Oxford's Global Priorities Institute, Open Philanthropy, and other EA-aligned organizations.

Their writing spans considerations of how much to explore and experiment in one's career, just how big of a difference there is in impact across careers, and reviews and guides for specific career paths that they think can make a large positive difference in the world.

You can look into their popular and informative podcast (which interviews people working on global challenges), their key ideas and guide, and their current list of especially pressing problems.

Core Materials (From 'Cause Prioritisation'):

- How not to waste your career by working on things that don't change the world (6 min video)
- This is your most important decision - 80,000 Hours (20 mins.)
- Prospecting for Gold (60 mins. or 30 mins. at 2x speed; we recommend following along with the transcript)
- How to compare different global problems in terms of impact (Article, 20 min)

Recommended reading

- Important Between-Cause Considerations - EA Forum (15 min)
- Which cause is most effective? (Article, 40 min)
- Cause Prioritisation flow chart from the Global Priorities Project
- The EA Career Planning Intensive - *A series of three five-week modules for planning your career to maximize your impact. In particular, see the* Reading List and Resources for Track I: Exploring Cause Areas for Maximum Career Impact.

Discussions and criticisms of the Scale, Neglectedness, Tractability framework:

- An argument that scale is often misused
- An argument that neglectedness is not always important
- An argument for adding "urgency"
- The ITN framework, cost-effectiveness, and cause prioritisation - criticism of the ITN framework from John Halstead.
- Understanding and evaluating EA's cause prioritisation methodology
- Doing good while clueless
- Factors other than ITN (Question on the forum)

More to explore

- EA Forum post, "Three Levels of Cause Prioritisation"
- The Global Priorities Institute's Research Agenda
- GiveWell on "Strategic Cause Selection"
- Implementing cause prioritisation at OpenPhil (Video, 23 min)
- Update on Cause Prioritization at Open Philanthropy

Exercise

For the exercise this week, take some time to reflect on the past weeks and what the most important considerations are for improving the world the most you can.

Reflecting back

You’ve covered a lot over the past weeks! We hope you found it an interesting and enjoyable experience. There are lots of major considerations to take into account in trying to do the most good you can, and lots of ideas may have been new and unfamiliar to you. This week we’d like you to reflect back on the Program with a skeptical and curious mindset.

We think it’s important to think hard about what the material you’ve covered implies about which things you should focus on going forward and how you prioritize your altruistic efforts.

To recapitulate what we’ve covered:

What topics or ideas from the program do you most feel like you don’t understand?

What seems most confusing to you about each one?

What idea from the program did you find surprising when you learned about it, and which you now think more or less makes sense and is important?

If you assumed this idea was true, what would the implications for your life and what you should work on be?

If you assumed this idea was false, what would the implications for your life and what you should work on be?

What idea from the program did you find surprising when you learnt about it, and think probably isn't right, or have reservations about?

If you assumed this idea was true, what would the implications for your life and what you should work on be?

If you assumed this idea was false, what would the implications for your life and what you should work on be?

Week 8: Where from Here? Planning a Career

"[I]f it's possible to find an option that's 100 times higher impact than your current best guess, then ten years in that path would achieve what could have otherwise taken people like you 1,000 years."
— The 80,000 Hours Team

One of the main ways in which we can affect the world for the better is through our careers. For this final week we hope to help you apply the principles of effective altruism to your own life and also critically reflect back on the rest of the program. Then, we consider how you can get more involved in the effective altruism community, should you wish to.

Required Materials

- A guide to using your career to solve the world's most pressing problems - 80,000 Hours (1 hour* - Read the sections that seem most relevant)
  - We also recommend you read a career or cause area review based on your interests
- About Us - Giving What We Can (1 min.)
- Effective altruism as one of the most exciting causes in the world (3 mins.)

Recommended reading

- A (free) weekly career planning course for positive impact - 80,000 Hours (8 weeks)
- Advice on how to read our advice - 80,000 Hours (10 mins.)
- Ideas for high-impact careers beyond our priority paths - 80,000 Hours (20 mins.)
- Problem areas beyond 80,000 Hours' current priorities - EA Forum (20 mins.)
- You have more than one goal, and that's fine (5 mins.)

More to explore

- Evidence-based advice on how to be successful in any job - 80,000 Hours (45 mins.)
- 'I give away half to three-quarters of my income every year' - The Guardian - *A lifestyle piece about one of Giving What We Can's members. You might also want to read these member profiles from Giving What We Can -* Jo, Arvind, and Catherine (5 mins.)

Exercise

Looking Forward

People often encounter interesting, important, or otherwise resonant ideas but, for a variety of reasons, don't end up continuing to explore them. If you found the ideas in this program worthwhile, it's worth making a plan for pursuing them further and creating accountability for yourself.

One way to do this is to go off and explore important global problems to resolve uncertainties you might have about prioritizing them. For example, you might be interested to know what experts project as the future of factory farming--if and when it might end, and what that might require.

For three potential problems you might work on with your career, identify a concrete question you could plausibly answer that would help you decide how important it is for you to work on that issue.

1. *First Problem Area: ___*
2. *Second Problem Area: ___*
3. *Third Problem Area: ___*

Another is to look into concrete career paths. You might look to the 80,000 Hours Five career categories for generating options, select three that seem most interesting for you to explore more, and ask your facilitator or the director of your program if they know people who you might speak with to learn more!

What are three potential career paths you'd be especially excited to learn more about, and why?

Finally, you could get involved with your local EA group and get to know other people interested in the principles covered in this program. Many groups run further programs, including in-depth programs that go into more detail on what you've learned here and career planning programs that help you chart a course toward a highly-impactful career.

What would you be most excited about your local EA group doing in the future? What are you most looking for in a group of people trying to improve the world?

Who should you contact to learn more about your local group and become more involved? You might ask your facilitator or look on your group's website.

Of course, if after this program you decide that EA is not for you, that’s ok as well. We hope the experience has been fruitful!