Introduction to Effective Altruism Program Syllabus

by the Centre for Effective Altruism, run by Effective Altruism Hungary

About the program

What the program is about[1]

Effective altruism (EA) is an ongoing project to find the best ways to do good, and put them into practice.

Our core goal with this program is to introduce you to some of the principles and thinking tools behind effective altruism. We hope that these tools can help you as you think through how you can best help the world.

We also want to share some of the arguments for working on specific problems, like global health or biosecurity. People involved in effective altruism tend to agree that, partly due to uncertainty about which cause is best, we should split our resources between problems. But they don't agree on what that split should be. **People in the effective altruism community actively discuss and disagree about which causes to prioritize and how**, even though we've learned a lot over the last decade. We hope that you will take these ideas seriously and think for yourself about which ways to help are most effective.

Finally, we give you some time at the end of the program to begin to reflect on how you personally can help to solve these problems. We don't expect you'll have an answer by the end of the sessions, but we hope you're better prepared to explore this further.

What the program involves

Each session of the program has a section of in-session materials and sometimes an exercise, which you will read with your peers and mentor during the 2-hour sessions. (Overall a 2-hour session will involve 1 hour of reading and one hour of discussions).

[1] Staff from the Centre for Effective Altruism drew up a draft of this curriculum, and then incorporated feedback from community members, subject matter experts, and program facilitators. Our goal is to introduce people to some of the core principles of effective altruism, to share the arguments for different problems that people in effective altruism work on, and to encourage you to think about what you want to do on the basis of those ideas. We also tried to give a balance of materials that is in line with the (significant) diversity of views on these topics within effective altruism. We think that these readings are interesting and give a good introduction, but we hope that you engage with them critically and not just take them all at face value. Ultimately we had to make many judgement calls, and other people would have drawn up a different curriculum. Once you've read this curriculum, we encourage you to explore other EA writings (e.g. on this wiki).

Beyond the in-session readings, there are more materials each session in ‘More to Explore’ — these are all optional and explore the themes of the session in more depth and breadth.

How we hope you’ll approach the program[2]

Taking ideas seriously.
Often, conversations about ideas are recreational: we enjoy batting around interesting thoughts and saying smart things, and then go back to doing whatever we were already doing in our lives. This is a fine thing to do — but at least sometimes, we think we should be asking ourselves questions like:

- “How could I tell if this idea was true?”
- “What evidence would it take to convince me that I was wrong about an idea?”
- “If it is true, what does that imply I should be doing differently in my life? What else does it imply I’m wrong about?”
- “How might this impact my plans for my career/life?”

And, zooming out:

- “Where are my blind spots?”
- “Which important questions should I be thinking about that I’m not?”
- “Do I really know if this idea/plan will help make things better or not?”

Answering these questions can help make our worldviews as accurate and full as possible and, by extension, help us make better decisions about things that we care about.

Disagreements are useful. When thoughtful people with access to the same information reach very different conclusions from each other, we should be curious about why and we should actively encourage people to voice and investigate where those disagreements are coming from. 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. If you start off disagreeing with someone then change your mind, that can be hard to admit, but we think that should be celebrated. Helping conversations become clearer by changing your mind in response to arguments you find compelling will help the community act to save lives more effectively Even if you don’t expect to end up agreeing with the other person, you’ll learn more if you acknowledge that you disagree and try to *understand* exactly how and why their views disagree with yours.

Be aware of our privilege and the seriousness of these issues. We shouldn’t lose sight of our privilege in being able to discuss these ideas, or that we are talking about real lives. We’re lucky to be in a position where we can have such a large impact, and this opportunity for impact is the consequence of a profoundly unequal world. Also, be conscious of the fact that people in this program come to these discussions with different ideas, backgrounds, and knowledge. 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 we may have personal ties to some of these areas.

Explore further. This fellowship aims to introduce people to effective altruism in a structured manner. There are far too many relevant topics, ideas, and research for all but a small fraction of them to fit into this very short program. If you are interested in these topics, you may find it very useful to dive into the linked websites, and the websites those sites link to, and so on.

[2] Inspired by Julia Galef’s Update Project

1) The effectiveness mindset

"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, explaining the need to prioritize given our limited resources.

If you want to use your time or money to help others, you probably want to help as many people as you can. But you only have so much time to help, so you can have a much bigger impact if you focus on the interventions that help more people rather than fewer.

But finding such interventions is incredibly difficult: it requires a "scout mindset" - seeking the truth, rather than to defend our current ideas.

Key concepts from this session include:

- **Scope sensitivity:** saving ten lives is more important than saving one, and saving a billion lives is a lot more important than saving ten.
- **Tradeoffs:** Because we have limited time and money, we need to prioritize between different ways to improve the world.
- **Scout mindset:** We'll be better able to help others if we're working together to think clearly and orient towards finding the truth, rather than trying to defend our own ideas. Humans naturally aren't great at this (aside from wanting to defend our own ideas, we have a host of other biases), but if we want to really understand the world, it's worth seeking the truth and trying to become clearer thinkers.

Optional material(s) before the session:

- Introduction to EA | Ajeya Cotra | EAGxBerkeley 2016 (Video - 30 mins.)
- Our top 3 lessons on how not to waste your career on things that don't change the world by 80,000 Hours (7 mins.)
- See "More to explore" for more!

In-session readings:

1) On effective altruism:
- Introduction to Effective Altruism (15 mins.)
- Extra - Four Ideas You Already Agree With (5 mins.)
- Extra - The world is much better; The world is awful; The world can be much better (5 mins.)

2) On scope sensitivity and tradeoffs:
- On Caring (10 mins)
- We are in triage every second of every day (5 mins.)

- Extra - Scope insensitivity: failing to appreciate the numbers of those who need our help (5 mins.)
- Extra - Purchase fuzzies and utilons separately (6 mins.)

3) On scout mindset and thinking clearly:
- Video to watch together -Why you think you're right – even when you're wrong (11 mins.)
- And read the tables from What cognitive biases feel like from the inside, or the full text (full text is extra)

4) Ending, if you have time
- 500 Million, But Not a Single One More (2 mins.)

More to explore

Other introductions

- *Doing Good Better* - Introduction through to the end of Chapter 3 (50 mins.)
- Effective Altruism: An Introduction - 80,000 Hours - *Ten curated episodes from The 80,000 Hours Podcast. (Ten 1.5 hour - 4 hour podcasts)*
- Effective altruism as I see it (7 mins.)
- No matter your job, here's 3 evidence-based ways anyone can have a real impact - 80,000 Hours (20 mins.)
- Other-centered ethics and Harsanyi's Aggregation Theorem (107 min)

Essays on caring

- The value of a life - Minding Our Own Way - *Disentangling the difference between the value of a life and what it costs to save a life in our broken world.* (15 mins.)
- Excited altruism - GiveWell - *Where does our own passion and excitement fit in?* (10 mins.)

Tradeoffs

- Tradeoffs - *How can we balance our own needs with the needs of others?* (5 mins.)
- Famine, affluence, and morality (15 mins.) *Note that many people in effective altruism disagree about exactly how demanding these ideas are.*
- Sustainable motivation - *How can we stay motivated when facing massive problems* (24 min talk)

Thinking carefully

- Minimal trust investigations (18 mins.)
- Double crux: a strategy for mutual understanding (17 mins.)
- Outline of Julia Galef's "Scout Mindset" (20 mins.)
- Humans are not automatically strategic (5 mins.)
- Beware surprising and suspicious convergence (22 mins.)

2) Differences in impact

Around 700 million people still live in poverty, mostly in low-income countries. Efforts to help them - by policy reform, cash transfers, or provision of health services - can be incredibly effective.

Alongside investigating this issue, we also discuss how much more effective some interventions are than others, and we introduce a simple tool for estimating important figures.

Key concepts from this session include:

- **Differences in impact:** It appears that some of our options to help do many times more good than others. People generally don't appreciate this, and so miss out on significant opportunities to help.
- **Fermi estimates:** When you're trying to make a decision, it can be useful to make a rough calculation for which option is best. Even if there's a lot of uncertainty, this can give you a rough answer, and can tell you which things are most important to estimate next.

Optional material(s) before the session:

- Read or watch Prospecting for Gold (55 mins.)
- GiveWell's "Giving 101" guide (click through to "next" at the bottom of each page, about 15 mins., but feel free to explore the rest of the site).
- See "More to explore" for more!

In-session readings:

1) Differences in impact:

- Comparing charities: How big is the difference? (5 mins.)
- Until the "Challenges addressed" part The moral imperative towards cost-effectiveness, or the full text as extra (10 mins.)

2) Fermi estimations and Global health

- A brief explanation of Fermi estimation and an example Fermi estimate (2 mins.)
- Global economic inequality (Our World in Data, 15 mins.)
- Extra - Global health (Our World in Data, 20 mins.)

3) Classic and more recent EA strategies for addressing global poverty:

- Classic - Health in poor countries problem profile (10 mins.)
- Recent
    - Introducing the Lead Exposure Elimination Project (7 mins.)
    - Extra - Donating money, buying happiness, read background and summary

More to explore

GiveWell and Open Philanthropy

GiveWell and Open Philanthropy are sister organizations in the effective altruism community. Both seek to identify outstanding giving opportunities, but they use different criteria and processes.

GiveWell has an emphasis on evidence-backed organizations within the global health and wellbeing space, while Open Philanthropy also supports high-risk, high-reward work, as well as work that could take a long time to pay off, in a variety of cause areas. We think this illustrates interesting methodological differences between attempts to answer the question "How can we do the most good?".

- Our Criteria - GiveWell and Process for Identifying Top Charities - GiveWell (20 mins.)
- Alexander Berger on Global health and wellbeing (3 hours)
- Hits-based Giving - Open Philanthropy (45 mins.)
- South Asian Air Quality Cause Investigation (Open Philanthropy) (18 mins.)

Cost-effectiveness methodology

- Prospecting for Gold (55 mins.)
- Finding the best charity requires estimating the unknowable. Here's how GiveWell does it. (Podcast - 1 hour 45 mins.)
- Why we can't take expected value estimates literally (even when they're unbiased)
- 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.)
- 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 yourself* (30 mins.)
- A framework for comparing global problems in terms of expected impact - 80,000 Hours (25 mins.)
- Subjective Confidence Intervals - Animal Charity Evaluators (10 mins.)
- RCTs in Development economics, their critics and their evolution (18 mins.)
- *How to Measure Anything*, Chapter 1 and 2 (50 mins.)

Mental health rather than physical health?

- Donating money, buying happiness and Happiness for the whole household - (30 mins. between them) *A cost effectiveness analysis that suggests that psychotherapy may be 9 times more effective than cash transfers (and thus competitive with GiveWell's top charities).*
- Using Subjective Well-Being to Estimate the Moral Weights of Averting Deaths and Reducing Poverty (52 mins.) - *Argument that subjective well-being is a better metric for determining value than physical health or wealth.*

Other newer strategies for improving human wellbeing

- Wave is a startup that is now the largest mobile money service in Senegal. Some of Wave's founders and early employees worked on it because they believe that it's an extremely effective way to improve the world.
- Ben Kuhn, Wave's CTO makes the case for founding a startup that serves emerging markets generally being an effective way of improving people's lives.
- Many people in the developing world commit suicide by drinking pesticide. It seems that we can significantly reduce suicide rates if we ban the more dangerous sorts of pesticide.
- Charity Entrepreneurship has incubated a number of charities in this space, focused on interventions that they think could be highly effective. These include:
    - The Centre for Alcohol Policy Solutions
    - Canopie
    - Family Empowerment Media
    - Happier Lives Institute
    - Suvita
    - Fortify Health

Effective aid

- Growth and the case against randomista development - *An argument that research on and advocacy for economic growth in low- and middle-income countries is more cost-effective than the things funded by proponents of randomized controlled trials development.* (1 hour - if you're short on time, read Sections 1-3)
- 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.)

Criticisms of the use of cost-effectiveness estimates

- Evidence, cluelessness, and the long term - *Evidence covers only the more immediate effects of any intervention, and it's highly likely the vast majority of the value is thereby omitted from the calculation.* (30 mins.)
- Charity Cost-Effectiveness in an Uncertain World – Center on Long-Term Risk - *Another way to deal with prioritization under uncertainty is to focus on actions that seem likely to have generally positive effects across many scenarios, rather than focusing on clear, quantifiable metrics.* (30 mins.)
- 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 mins.)
- Why Charities Usually Don't Differ Astronomically in Expected Cost-Effectiveness - *An argument about how those in the effective altruism movement might overestimate the extent to which charities differ in their expected marginal cost-effectiveness.* (40 mins.)

3) Radical empathy

“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)

Should we care about non-human animals? We'll show how it can be important to care impartially, rather than ignoring weird topics or unusual beneficiaries.

We'll also cover expected value theory (which helps when we're uncertain about the impact of an intervention), and give some ideas for how we could improve the lives of animals that suffer in factory farms.

Key concepts from this session include:

- **Impartiality:** helping those that need it the most, only discounting people according to location, time, and species if those factors are in fact morally relevant.
- **Expected value:** We’re often uncertain about how much something will help. In such circumstances, it may make sense to weigh each of the outcomes by the likelihood that they occur and pick the action that looks best in expectation.
- **The importance (and difficulty) of considering unusual ideas:** Society’s consensus has been wrong about many things over history (e.g. the sun circling the Earth, the morality of slavery). In order to avoid making similar mistakes, we need to be open to considering unusual ideas and moral positions, while still thinking critically about the issues and acting cooperatively with others.

Optional material(s) before the session:

- See exercise below the readings (10 mins)
- On "fringe" ideas (7 mins.)
- The possibility of an ongoing moral catastrophe (summary) (8 mins.)
- How Students Will Lead the Alternative Protein Revolution - Video (26 mins.)
- Want to help animals? Focus on corporate decisions, not people’s plates. (10 mins.)
- Animal Advocacy Careers (website, explore)
- See “More to explore” for more!

In-session readings:

1) Impartiality and radical empathy:

- Radical Empathy (10 mins)
- Moral Progress and Cause X (5 mins.)

2) Expected value calculation

- Expected Value (5 mins.)
- Hits based giving (15 mins.)

3) The case for caring about animal welfare:

- All Animals Are Equal - *An excerpt from the opening chapter of* Animal Liberation *(1975), widely regarded as the founding text of the animal rights movement.* (5 mins.)
- Animal Welfare Cause report, Founders Pledge (10 mins.)

Exercise (10 mins)

This session's exercise is 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.

A letter to the past (10 mins.)

This exercise asks you to explore what it would take to change your mind about something important.

Imagine someone from the past who 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 have been friends. Unfortunately, many 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 short letter to this historical friend arguing that they should care about 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 realize 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

More to explore - An expanding moral circle?

- *The Expanding Circle* pg. 111-124 'Expanding the Circle of Ethics' section (20 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.)
- Our descendants will probably see us as moral monsters. What should we do about that? - 80,000 Hours - *A conversation with Professor Will MacAskill.* (Podcast - 1 hour 50 mins.)
- The Possibility of an Ongoing Moral Catastrophe (full text of the required article, 30 mins.)

The case for caring about animal welfare

- The Case Against Speciesism - Centre for Reducing Suffering (10 mins.)
- Factory Farming - 80,000 Hours (5 mins.)
- Should animals, plants, and robots have the same rights as you? - Vox (20 mins.)
- *Animal Liberation,* Chapter 3 - Down on the factory farm (1 hour.)
- 2017 Report on Consciousness and Moral Patienthood - An *investigation into what types of beings merit moral concern.* (6 hours, skimmable)
- Suffering in Animals vs. Humans (13 mins.)

Reforming animal agriculture

- *Dominion* - *Dominion uses drones, hidden and handheld cameras to expose the dark side of modern animal agriculture.* (Film - 2 hours)
    - 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.
- Food impacts *a tool to explore the moral impact of different dietary choices.*
- A New Agricultural Revolution (~22 mins. and transcript available; Q&A after Friedrich's talk is optional)

Wild animal welfare

- Wild animal suffering: An introduction - Animal Ethics - *An argument for us to take into account the wellbeing of animals that live in the wild.* (10 mins)
- Wild Animal Initiative's FAQ

Criticism of EA-related animal welfare work

- How the animal movement could do even more good (11 mins.)
- EAA is relatively overinvesting in corporate welfare reforms *There is also an interesting response to this post from Saulius in the first set of comments.* (7 mins.)
- Against the Moral Standing of Animals *critique of arguments that animals deserve moral standing*
- What's Wrong with Speciesism? *Argument that animals do deserve moral standing, but lower moral status*

4) Our final century?

“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

Humanity appears to face existential risks: a chance that we'll destroy our long-term potential. We’ll examine why existential risks might be a moral priority, and explore why they are so neglected by society. We’ll also look into one of the major risks that we might face: a human-made pandemic, worse than COVID-19.

Alongside this we'll introduce you to the concepts of neglectedness, marginal thinking, and explore whether you could lose all of your impact by missing one crucial consideration.

We’ll also introduce the following concepts:

- **The importance, neglectedness, tractability framework:** The most important problems generally affect a lot of people, are relatively under-invested in, and can be meaningfully improved with a small amount of work.
- **Thinking on the margin:** If you're donating $1, you should give that extra $1 to the intervention that can most cost-effectively improve the world. There are many great initiatives with a very high average impact per dollar that will have a low marginal impact because they can't get the same efficiency at scale (they display "diminishing marginal returns").
- **Crucial considerations:** It can be extremely hard to figure out whether some action helps your goal or causes harm, particularly if you’re trying to influence complex social systems or the long-term. This is part of why it can make sense to do a lot of analysis of interventions you’re considering.

Required Materials

Optional material(s) before the session:

- Chapter 2 of *The Precipice* (Book - 35 mins.)
- See “More to explore” for more!

1) Existential risks:

- The case for reducing existential risks until last section (25 mins.)
- (Extra) - read the last section on “Who shouldn’t prioritise safeguarding the future?”

2) Thinking on the margin and ITN framework:

- Marginal Impact (3 mins.)
- A framework for comparing global problems in terms of expected impact (15 mins.)
- (Extra) See summary of how 80,000 Hours apply this framework in their profile on climate change, or feel free to read on (20 mins.)

3) Risks from pandemics and improving biosecurity:

- Why experts are terrified of a human-made pandemic — and what we can do to stop it (15 mins.)
- Concrete Biosecurity Projects - (7 mins.)
- (Extra) - Biosecurity needs engineers and material scientists (4 mins.)

More to explore

Existential risks

- "Future risks" chapter of *The Precipice*, introduction and "Pandemics" section. Stop at "Unaligned artificial intelligence" (Book - 25 mins.)
- Crucial considerations and wise philanthropy (24 mins.)
- Policy and research ideas to reduce existential risk - 80,000 Hours (5 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.)

Criticism

- Democratizing Risk: In Search of a Methodology to Study Existential Risk (50 mins.)
- A critical review of "The Precipice" (maybe come back to the "Unaligned Artificial Intelligence" section next session, once you've engaged with the argument for AI risk).

Biosecurity

- Reducing Global Catastrophic Biological Risks Problem Profile - 80,000 Hours (1 hour)
- The Apollo Report
- Global Catastrophic Risks Chapter 20 - Biotechnology and Biosecurity Overview from ~15 years ago on how *biotechnological power is increasing exponentially, as measured by the time needed to synthesize a certain sequence of DNA. This has important implications for biosecurity.* (1 hour)
- Open until dangerous: the case for reforming research to reduce global catastrophic risk (Video - 50 mins.)
- Dr. Cassidy Nelson on the twelve best ways to stop the next pandemic (andlimit COVID-19) 80k podcast interview (podcast - 2.5 hours)
- Andy Weber on rendering bioweapons obsolete and ending the new nuclear arms race 80k podcast interview (podcast - 2 hours)

- Article on information hazards in biotechnology (15 mins.)
- Using Export Controls to Reduce Biorisk (6 mins.)
- Lynn Klotz on improving the Biological and Toxin Weapons Convention (BWC) (10 mins.)
- Horsepox synthesis: A case of the unilateralist's curse? (8 mins.)

Climate Change

- Climate Change Problem Profile - 80,000 Hours - *An analysis of the worst risks of climate change, some of the most promising ways to reduce them, and how to prioritize climate change against other problems.* (30 mins.)
- Effective Environmentalism (Website)

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.)
- Risks from Nuclear weapons - A series of posts exploring the extent to which nuclear risk reduction is a top priority, and the most effective ways to reduce nuclear risk.
- Nuclear security - Brief summary + relevant articles on the *EA Forum*

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.)
- Modeling Great Power conflict as an existential risk factor (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)

5) What could the future hold? And why care?

"Longtermism" is the view that improving the long term future is a key moral priority of our time. This can bolster arguments for working on reducing some of the extinction risks that we covered in the last section.

We'll also explore some views on what our future could look like, and why it might be pretty different from the present. And we'll introduce forecasting: a set of methods for improving and learning from our attempts to predict the future.

Key concepts from this session include:

- **Impartiality:** helping those that need it the most, only discounting people according to location, time, and species if those factors are in fact morally relevant.
- **Forecasting:** Predicting the future is hard, but it can be worth doing in order to make our predictions more explicit and learn from our mistakes.

You will also practice the skill of **calibration**, with the hope that when you say that something is 60% likely, it will happen about 60% of the time. This is important for making good judgments under uncertainty.

Pre-session readings

- This Can't Go On (15 mins.)
- Calibrate Your Judgment app (30 minutes)

In-session readings

1) Hinge of History:

- All Possible Views About Humanity's Future Are Wild (15 mins.)

2) The case for and against longtermism:

- What We Owe the Future, Chapter 1 (22 mins)
- (Extra) Why I Find Longtermism Hard, and How to Deal with That Difficulty (10 mins.)
- (Extra) Why I am probably not a longtermist (11 mins.)

3) To what extent can we predict the future? How?

- Superforecasting in a nutshell (3 mins.)
- Longtermism and animal advocacy (3 mins.)
- Top open Metaculus forecasts (5-10 mins) Read the first few dozen results and reflect on what you find important and surprising. These are average predictions of

how several important trends will unfold over the coming years. We're not sure how accurate they'll be, but we think it gives a glimpse into the future.

More to explore

Global historical trends

- How big a deal was the Industrial Revolution? (1 hour 20 mins.)
- Three wild speculations from amateur quantitative macrohistory (10 mins.)
- Modeling the Human Trajectory - Open Philanthropy Project (30 mins.)

Forecasting

- Efforts to Improve Accuracy in our Judgements and Forecasts - Open Philanthropy (10 mins.)
- How accurate are Open Phil's predictions? (18 mins.)
- Efforts to Improve the Accuracy of Our Judgments and Forecasts - *exploring the ways to work on and importance of improving our calibration*

The case for longtermism

- Orienting towards the long-term future (Video - 25 mins.)
- The Case for Strong Longtermism - Global Priorities Institute (1 hour 20 mins.)
- The epistemic challenge to longtermism (2 min. Discussion of a somewhat longer paper)
- *The Precipice,* Appendix B - Population Ethics and Existential Risk (10 mins.)
- 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.)
- 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.)
- Major UN report discusses existential risk and future generations (summary) (18 mins.)

Criticism of longtermism

- Against Longtermism (1 min. summary, 5. min read)
- This short comment, and the more academic pieces that it links to (2 mins. for the comment, lots to explore)
- How the simulation argument dampens future fanaticism
- How much current animal suffering does longtermism let us ignore? (9 mins.)
- This comment, with ten reasons to work on near-term causes (2 mins.)

Suffering risks

- S-risks: Why they are the worst existential risks, and how to prevent them (20 mins.)

6) Risks from artificial intelligence (AI)

Transformative artificial intelligence may well be developed this century. If it is, it may begin to make many significant decisions for us, and rapidly accelerate changes like economic growth. Are we set up to deal with this new technology safely?

This session you'll also:

- Learn about **Bayes' rule**, a guide to how to change your beliefs as you get new evidence.

Optional pre-session readings

- Why AI alignment could be hard (18 mins)
- Preventing an AI-related catastrophe (60 mins)
- AI Timelines: Where the arguments and the "Experts," stand (13 mins.)

In-session readings:

1) The case for worrying about risks from artificial intelligence:

- The case for taking AI seriously as a threat to humanity (30 mins.)
    - read until section 8 or further as extra reading

2) Strategies for reducing risks from unaligned artificial intelligence:

- Read on or the other depending on your background:
    - The longtermist AI governance landscape: a basic overview (14 mins.)
    - AI Safety researcher career review (10 mins.)
- (extra) Preventing an AI-related catastrophe (read the "What you can do concretely to help" section)

3) Bayes' rule and evidence:

- Bayes' rule: Guide (15 mins. for the short version)
- Making beliefs pay rent (5 mins.)
- (Extra) What is evidence? (4 mins.)

4) Suffering risks:

- S-risks: Why they are the worst existential risks, and how to prevent them (13 mins.)

More to explore

The development of artificial intelligence

- 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 - 1 hour 30 mins.)
- The Artificial Intelligence Revolution: Part 1 - *A fun and interesting exploration of artificial intelligence by the popular blogger Tim Urban.* (45 mins.)

Other resources on aligning artificial intelligence

- AGI Safety Fundamentals Curricula
- My personal cruxes for working on AI safety (65 mins.)
- Professor Stuart Russell on the flaws that make today's AI architecture unsafe & a new approach that could fix it (Podcast - 2 hours 15 mins.)
- 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.* (1 hour)
- *The Precipice* - Chapter 5 (pages 138-152) - Unaligned Artificial Intelligence (25 mins.)
- What Failure Looks Like (12 mins.) - *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*
- AGI Safety from first principles (1 hour 15 mins.) - *one AI researcher's take on the specific factors for the problem of aligning general AI*
- *Human Compatible: Artificial Intelligence and The Problem of Control* (Book)
- *The Alignment Problem: Machine Learning and Human Values* (Book)

Governance for artificial intelligence

- The new 30-person research team in DC investigating how emerging technologies could affect national security - 80,000 Hours - *How might international security be altered if the impact of machine learning is similar in scope to that of electricity?* (Podcast - 2 hours)
- Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority - Center 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.)

Technical AI alignment work

- AI Alignment Landscape (Video - 30 mins.)
- AI safety starter pack (7 mins.)
- How to pursue a career in technical AI alignment (59 mins.)

- Technical Alignment Curriculum (readings for a 7 week course)
- The Alignment Forum, especially their core sequences

Criticisms of worries about AI risk

- How sure are we about this AI stuff? (26 mins.)
- A tale of 2.75 orthogonality theses (20 mins.)
- How to know if AI is about to destroy civilization (summary, 2 mins.)
- The AI Messiah (and the first comment) (5 mins.)
- How good is humanity at coordination? (4 mins.)

7) What do you think?

"It is one of the unfortunate truisms of the human condition that there is hardly a good idea, noble impulse, or sound suggestion that can't be (and isn't eventually) adopted and bastardized by zealots… One iteration of this tendency is in the idea of "effective altruism."
- K. Berger & R. M. Penna

This session, we'll give you time to reflect on what you think of effective altruism, and of the specific potential priorities you've heard about so far.

We are dedicating a session to this because, to whatever extent we are wrong, realizing and correcting our mistakes will allow us to do more good. Honestly reckoning with strong counterarguments (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, GiveWell polled a sample of people demographically similar to recipients of programs it supports 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.

A key concept for this session is the importance of **forming independent impressions.** In the long run, you're likely to gain a deeper understanding of important issues if you think through the arguments for yourself. But (since you can't reason through everything) it can still sometimes make sense to defer to others when you're making decisions.

Pre-Session materials:

Independent impressions:

- Independent impressions - (2 mins.)

Recent critique of effective altruism. Read articles in More to Explore for others:

- Notes on Effective Altruism (20 mins.)

While we've covered some of the most popular EA causes above, there are many other causes that we haven't had space to cover. Please skim over this list of other causes to get a sense of other ideas that people in EA have discussed. (Note, you don't need to read this whole post in detail!)

Exercise, which includes reading and reflecting on criticisms of ideas covered in previous sessions (see below).

In-session Materials

For the exercise this session, we will take some time to reflect on the ideas we've engaged with over the past sessions. Our goal is to take stock and to identify our concerns and uncertainties about EA ideas.

1) What are your concerns about EA? (5-10 mins.)

We've covered a lot: the philosophical foundations of effective altruism, how to compare causes and allocate resources, and a look at 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, or anything you'd like!

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

Reflecting back (15-20 mins.)

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

To recapitulate what we've covered:

The Effectiveness mindset

Over the course of sessions 1 and 2, we aim to introduce you to the core principles of effective altruism. We use global health interventions, which has been a key focus area for effective altruism, to illustrate these principles, partly because we have unusually good data for this cause area.

Differences in impact

We continue to explore the core principles of effective altruism, particularly through the lens of global health interventions because they are especially concrete and well-studied. We focus on giving you tools to quantify and evaluate how much good an intervention can achieve; introduce expected value reasoning; and investigate differences in expected cost-effectiveness between interventions.

Radical empathy

The next section focuses on your own values and their practical implications. We explore who our moral consideration should include. We focus especially on farmed animals as an important example of this question this session.

Our final century?

This session we'll focus on existential risks: risks that threaten the destruction of humanity's long-term potential. We'll examine why existential risks might be a moral priority, and explore why existential risks are so neglected by society. We'll also look into one of the major risks that we might face: a human-made pandemic, worse than COVID-19.

What could the future hold? And why care?

This session we explore what the future might be like, and why it might matter. We'll explore arguments for "longtermism" - the view that improving the long term future is a key moral priority. This can bolster arguments for working on reducing some of the extinction risks that we covered in the last two sessions. We'll also explore some views on what our future could look like, and why it might be pretty different from the present.

Risks from artificial intelligence

Transformative artificial intelligence may well be developed this century. If it is, it may begin to make many significant decisions for us, and rapidly accelerate changes like economic growth. Are we set up to deal with this new technology safely?

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?

List one idea from the program that you found surprising at first, and which you now think more or less makes sense and is important?

How could this idea be wrong? What's the strongest case against it?

List one idea from the program that you found surprising at first, and think probably isn't right, or have reservations about.

*What's the strongest case **for** this idea?*

What are your key hesitations about that case?

More to explore

Effectiveness is a conjunction of multipliers (5 mins.) - one take on why it matters so much to think carefully and critically about which of the above perspectives is right.

Types of criticism

- Disagreeing about what's effective isn't disagreeing with effective altruism - *Rob Wiblin differentiates critiques of effective altruism as a concept and critiques of the ways EAs attempt to apply this concept.* (5 mins.)
- Four categories of effective altruism critiques (4 mins.)

Systemic change

- Response to Effective Altruism, Iason Gabriel (1 min.)
- Effective altruists love systemic change - *Robert Wiblin argues why EA does not, in fact, neglect systemic change.* (13 mins.)
- Beware Systemic Change (15 mins.)
    - Critique of pursuing systemic change. How hard is it to figure out what systemic changes will make things better?
    - This is partly an expression of disagreement with others in EA who have embraced systemic change, which was itself partly a response to criticisms like those in the Boston Review

Is effective altruism a question or an ideology, or both?

- Effective Altruism is a Question (not an ideology) (5 mins.)
- Effective Altruism is an Ideology, not (just) a Question (24 mins.)

General criticisms of effective altruism

- Notes on Effective Altruism (20 mins.)

- The Centre for Effective Altruism's responses to some common objections (10 mins.)
- Responses to The Logic of Effective Altruism (~20 mins., pick a few to read) Note that these critiques are from 2015.
    - Recommended excerpts
        - Daron Acemoglu
        - Angus Deaton
        - Jennifer Rubenstein
        - Iason Gabriel
        - Peter Singer's response
    - How to view these: click the names under "Responses" at the bottom of the original article
- Towards Ineffective Altruism (15 mins.)
- A critique of effective altruism (11 mins.)
- Another Critique of Effective Altruism (5 mins.)
- The motivated reasoning critique of effective altruism (34 mins.)
- 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 summarizes the problem and suggests ways to resolve such issues.* (16 mins.)
- 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 mins.)
- Free-spending EA might be a big problem for optics and epistemics (12 mins.) - *EA forum post on risks associated with EA spending trends*
- 80,000 hours' anonymous flaws in EA
- Critiques of EA that I want to read (16 mins)
- Effective Altruism: Not Effective and Not Altruistic (27 mins.)
- Stop the Robot Apocalypse - Amia Srinivasan - (15 mins.)
- EA and the current funding situation - *not exactly criticism, but discusses some potential pitfalls of EA's current funding situation* (35 mins.)

Deference and forming inside views

- Some thoughts on deference and inside view models (14 mins.)
- A sketch of good communication (4 mins.)
- How I formed my own views on AI safety (21 mins.)
- Deference Culture in EA (8 mins.)
- Bad Omens in Current Community Building (27 mins.)

Criticism of EA methods

- A philosophical review of Open Philanthropy's Cause Prioritisation Framework (42 mins.)
- Evidence, Cluelessness, and the Long Term - Hilary Greaves - (30 mins.)
- 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)

- 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)
- Ethical Systems - *Check out other ethical systems not discussed yet in the program. Which ones resonate most with you?* (Varies)
- Summary review of ITN critiques (8 mins.)

Criticism of EA principles

- Pascal's Mugging *Critique of the application of expected value theory. How do you deal with very low probability events that would be disastrous if they took place?* (5 mins.)
- Ethical Systems - *Check out other ethical systems not discussed yet in the program. Which ones resonate most with you?* (Varies)
- AI alignment, philosophical pluralism, and the relevance of non-Western philosophy - *Short talk* (18 mins.)
- The Repugnant Conclusion - *Total utilitarianism (maximizing overall wellbeing) implies that it's better to have many many beings with infinitesimally positive wellbeing to a smaller number of beings that are all extremely well off. Some people find this counterintuitive, but there's significant debate on this.* (Video - 6 mins.)
- Utility monster - *Another thought experiment suggesting that trying to maximize wellbeing may have counterintuitive implications* (5 mins.)
- The bullet-swallowers - *Scott Aaronson describes how some theories (like EA) force you to either swallow some tough conclusions or dodge them by contorting the theory.* (2 mins.)

8) Putting it into practice

In this final section, we hope to help you apply the principles of effective altruism to your own life and career.

You probably won't be ready to make a career change just yet - you might want to read and reflect more before you do that. So instead we'll help you to think through some of your key uncertainties, generate tests for those uncertainties, and plan out how you can make sure you follow through on your intentions.

Optional pre-session readings

- My current impressions on career choice for longtermists[3] (40 mins., but we encourage you to skim and focus on the sections you're most interested in)
- Giving What We Can (5-10 mins exploring the site)
- Givewell (5 mins. exploring the site)

Required Materials

1) Two different attitudes that you can try on this session:
- Call to vigilance (5 mins.)
- Effective altruism as one of the most exciting causes in the world (3 mins.)

2) Career choice and donations:
- Summary of 80,000 Hours' key ideas (5 mins.)
- Exercise (see below)

3) Dealing with demandingness:
- You have more than one goal, and that's fine (5 mins.)

[3] Holden notes that this may extrapolate to EA more broadly, but he wanted to focus on the parts of EA that he was more familiar with (longtermism).

Exercise

For this exercise, you'll begin to think about what these ideas might mean for your life. We don't expect you to come up with a full plan now: you should probably continue to explore the ideas and think through various options. But we think that it can be useful to come up with some initial guesses at this stage, to structure your thinking.

Based on the readings in previous sessions, which global problems do you think are most pressing and why? (Remember, experts are quite uncertain about this question!)

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

What are your 3-5 biggest uncertainties about the above?

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

What could you do over the next few weeks to explore those uncertainties? (For example, do more reading, talk things through with a friend, or write an EA Forum post on your key uncertainties to get feedback.)

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

What aptitudes are you most interested in exploring or using next? You might want to think about what you're unusually good at, what activities make you feel energized, and what skills seem especially useful for addressing the problems you listed above.

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

How could you begin to test out those aptitudes over the next few weeks?

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

While you're figuring out your uncertainties, are there any actions to improve the world that you want to do now? (E.g. make a donation, or do things that make it more likely that you remember to apply these ideas in your life (like signing up to newsletters, committing to discuss your career with a friend, or making a task to apply for high-impact jobs/internships)).

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

More to explore

Career advice

- Evidence-based advice on how to be successful in any job - 80,000 Hours (45 mins.)
- A (free) weekly career planning course for positive impact - 80,000 Hours (8 weeks)
- Probably Good - *a slightly different perspective on effective career choice.*
- Magnify Mentoring
- 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.)
- Why founding charities is one of the highest impact things you can do (5 mins.)
- Animal Advocacy Careers

Cross-cause career profiles

- Founder of new projects tackling top problems career profile (10 mins.)
- Operations management in high-impact organizations career profile (10 mins.)

Donations

- A guide to giving effectively (11 min article or 17 min video)
- Why make a giving pledge when you could just donate? (5 min)
- Media coverage about people giving effectively: "'I give away half to three-quarters of my income every year"' - The Guardian and "New year's resolutions: 'I'm going to give away 10% of my income" - The Guardian - *lifestyle pieces about Giving What We Can's members.*
- *Written member profiles* or videos of Giving What We Can members sharing their experiences and motivations

Volunteering

- Where's the best place to volunteer? - 80,000 Hours (3 mins.)
- Is skilled volunteering for you? - Animal Advocacy Careers (2 mins.)
- Paths to Impact for EA Working Professionals - High Impact Professionals (3 mins.)

Other

- Take action - EA Forum
- Doing good together (40 mins.)
- Rowing, Steering, Anchoring, Equity, Mutiny *Holden Karnofsky explores some key different approaches to improving the world.* (22 mins.)