The Centre for Effective Altruism's

Effective Altruism Handbook

2nd Edition

Dear Reader,

This guide is an introduction to some of the core concepts in effective altruism. If you're new to effective altruism, they should give you an overview of the key ideas and problems that we've found so far. But if you have already engaged with effective altruism, there should be some new insights amongst the familiar ideas.

The pieces are a mix of essays and adaptations of conference talks. We've tried to put them in an order that makes sense. Together we think they cover some of the key ideas in the effective altruism community.

These essays were selected by the Centre for Effective Altruism. Of course, distilling a community's ideas requires making some difficult judgement calls, and we had to exclude lots of important work. Others in the community would have made a different selection, so at the end we link to other resources that you might find helpful. We are very grateful to all of the authors, for giving us permission to include their work.

Happy reading!

P.S. If you have any comments, please email content@effectivealtruism.org.

These resources can also be accessed online.

Contents

PRINCIPLES

THE LONG-TERM FUTURE

PROMISING CAUSES

FURTHER RESOURCES

Principles

How should we work out how to do good?

Introduction to Effective Altruism

June, 2016

An outstanding opportunity to do good

History contains many examples of people who have had a huge positive impact on the world.

Irena Sendler saved 2,500 Jewish children from the Holocaust by providing them with false identity documents and smuggling them out of the Warsaw ghetto. Norman Borlaug's research into disease-resistant wheat precipitated the 'Green Revolution'; he has been credited with saving hundreds of millions of lives. Stanislav Petrov prevented all-out nuclear war simply by staying calm under pressure and being willing to disobey orders.

These people might seem like unrelatable heroes, who were enormously brave, or skilled, or who just happened to be in the right place at the right time. But many people can also have a tremendous positive impact on the world, if they choose wisely.

This is such an astonishing fact that it's hard to appreciate. Imagine if, one day, you see a burning building with a small child inside. You run into the blaze, pick up the child, and carry them to safety. You would be a hero. Now imagine that this happened to you every two years - you'd save dozens of lives over the course of your career. This sounds like an odd world.

But current evidence suggests that this is the world that many people live in. If you earn the typical income in the US, and donate 10% of your earnings each year to the Against Malaria Foundation, you will probably save dozens of lives over your lifetime.

In fact, the world appears to be even stranger. Donations aren't the only way to help: many people have opportunities that look higher-impact than donating to global poverty charities. How? First, many talented people can have a greater impact by working directly on important problems than by donating. Second, other causes might prove even more important than global poverty and health, as we'll discuss below.

Many attempts to do good fail, but the best are exceptional

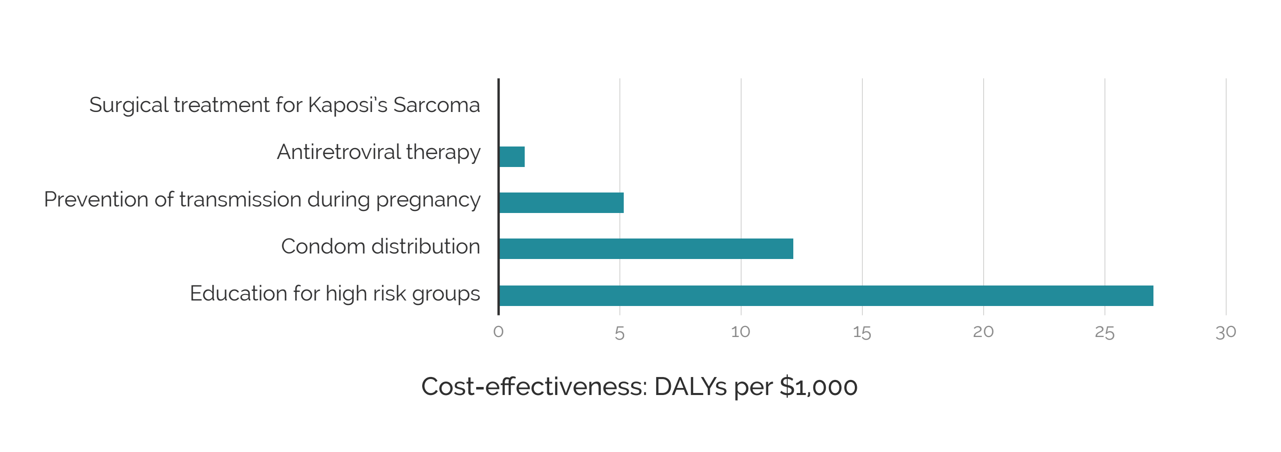
In most areas of life, we understand that it's important to base our decisions on evidence and reason rather than guesswork or gut instinct. When you buy a phone, you will read customer reviews to get the best deal. Certainly, you won't buy a phone which costs 1,000 times more than an identical model.

Yet we are not always so discerning when we work on global problems.

Below is a chart from an essay by Dr Toby Ord, showing the number of years of healthy life (measured using DALYs) you can save by donating $1,000 to a particular intervention to reduce the spread of HIV and

AIDS. The chart shows figures for five different strategies.
The first intervention, surgical treatment, can't even be seen on this scale, because it has such a small impact relative to other interventions. And the best strategy, educating high-risk groups, is estimated to be 1,400 times better than that. (It's possible that these estimates might be inaccurate, or might not capture all of the relevant effects. But it seems likely that there are still big differences between interventions.)

We suspect that the difference in intervention effectiveness is similarly large in other cause areas, though we don't have as clear data as we do in global health. Why do we think this? Partly because most projects (in many domains for which we have data) don't appear to have a significant positive impact. And, more optimistically, because there appear to be some interventions which have an enormous impact. But without knowing which experts to trust, or which techniques to trust in one's own research, it can be very hard to tell these apart.



Which interventions have the highest impact remains an important open question. Comparing different ways of doing good is difficult, both emotionally and practically. But these comparisons are vital to ensure we help others as much as we can.

It's important to work on the right problems

The media often focuses on negative stories.

But in many ways, the world is getting better. Concerted efforts to improve the world have already had phenomenal success. Let's consider just a few examples. The number of people living under the World Bank's poverty line has more than halved since 1990. The Cold War saw thousands of nuclear weapons trained across the Atlantic, but we survived without a single nuclear strike. Over the last few centuries, we have criminalized slavery, dramatically decreased the oppression of women, and, in many countries, done a great deal to secure the rights and acceptance of people who are gay, bi, trans or queer.

Nevertheless, many problems remain. Around 800 million people live on less than $2 per day. Climate change and disruptive new technologies have the potential to negatively impact billions of people in the future. Billions of animals, who may well have lives that matter, spend short lives of suffering in factory farms. There are so many problems that we need to think carefully about which ones we should prioritize solving.

The cause that you choose to work on is a big factor in how much good you can do. If you choose a cause where it's not possible to help very many people, or where there just aren't any good ways to solve the relevant problems, then you will significantly limit the amount of impact you can have.

If, on the other hand, you choose a cause with great prospects and tested solutions, you may have an enormous impact. For instance, some attempts to reduce the suffering of animals appear to be incredibly effective. By thinking carefully and acting strategically, a small group of campaigners – with limited budgets – have helped improve the conditions of hundreds of millions of chickens who were suffering in US factory farms.

Many people are motivated to do good, but already have a cause of choice before beginning research. There are lots of reasons for this, such as personal experience with a problem, or having a friend who's already raising money for a particular organization.

But if we choose a cause that simply happens to be salient to us, we may overlook the most important problems of our time. Given that most interventions seem to have low impact, we're likely to focus on something that is not very impactful if we don't pick carefully. But it may be even worse than this: issues that are salient to us are probably also salient to others like us, so it's likely there will be lots of other people working on those issues. This might mean that our additional efforts have even less impact. So we can do more good if we carefully consider many causes, rather than stopping at the first one we're drawn to.

By remaining open to working on different causes, we're able to change course to where we can make the biggest difference, without restricting ourselves too early.

Promising causes

How, then, can we figure out which causes we should focus on? Researchers have found the following framework to be useful. Working on a cause is likely to be high impact to the extent that it is:

- Great in *scale* (it affects many people's lives, by a great amount)
- Highly *neglected* (few other people are working on addressing the problem), and
- Highly *solvable* (additional resources will do a great deal to address it).

On the basis of this reasoning, there are several cause areas that appear particularly likely to be high-impact.

These choices are not immutable. They simply represent best guesses about where we can have the most impact, given the evidence currently available. As new evidence comes to light that suggests different causes are more promising, we should consider working on those instead. It's also worth keeping in mind that even if a person is motivated to choose a good cause rather than the best cause, their impact can still be much larger than it might have been.

We'll discuss three main areas. We start with the more intuitive area of global poverty, then turn to work to improve animal welfare. Finally, we look into less intuitive, but possibly more impactful, work to improve the long-term future.

Fighting extreme poverty

Diseases associated with extreme poverty, such as malaria and parasitic worms, kill millions of people every year. Also, poor nutrition in low-income countries can lead to cognitive impairment, birth defects and growth stunting.

Much of this suffering can be relatively easily prevented or mitigated. Antimalarial bednets cost around $2.50 each. GiveWell, an independent charity evaluator, estimates that they can significantly reduce malaria rates. Even simply transferring money to people who are very poor is a relatively cost-effective way of helping people.

Not only does improving health avert the direct suffering associated with sickness and death, it also allows people to participate more fully in education and work. Consequently, they earn more money, and have more opportunities later in life.

Animal suffering

The advent of industrialized agriculture means that billions of animals each year are kept in inhumane conditions on factory farms. Most have their lives ended prematurely when they are slaughtered for food. Advocates for their welfare argue that it is relatively cheap to reduce demand for factory farmed meat, or enact legislative changes that improve the welfare of farmed animals. Because of the huge numbers of animals involved, making progress on this issue could avert a very large amount of suffering.

Improving the long-term future

Most of us care not just about this generation, but also about preserving the planet for future generations. Because the future is so vast, the number of people who could exist in the future is probably many times greater than the number of people alive today. This suggests that it may be extremely important to ensure that life on earth continues, and that people in the future have positive lives. Of course, this idea might seem counterintuitive: we don't often think about the lives of our great-grandchildren, let alone their great-grandchildren. But just as we shouldn't ignore the plight of the global poor just because they live in a foreign country, we shouldn't ignore future generations just because they are less visible.

Unfortunately, there are many ways in which we might miss out on a very positive long-term future. Climate change and nuclear war are well-known threats to the long-term survival of our species. Many researchers believe that risks from emerging technologies, such as advanced artificial intelligence and designed pathogens, may be even more worrying. Of course, it is hard to be sure exactly how technologies will develop, or the impact they'll have. But it seems that these technologies have the potential to radically shape the course of progress over the centuries to come. Because of the scale of the future, it seems likely that work on this problem is even more high impact than work on the previous two cause areas.

Yet existential risks stemming from these technologies have been surprisingly neglected – there are just tens of people working on risks from AI or pathogens worldwide. US households spend around 2% of their budgets on personal insurance, on average. If we were to spend a comparable percentage of global resources on addressing risks to civilization, there would be millions of people working on these problems, with a budget of trillions of dollars per year. But instead, we spend just a tiny fraction of that amount, even though such risks may become substantial in the decades to come. If we value protection against unlikely but terrible outcomes individually, as our insurance coverage suggests we do, we should also value protection against terrible outcomes collectively. After all, a collective terrible outcome, like human extinction, is terrible for everyone individually, too. For this reason, it seems prudent for our civilization to spend more time and money mitigating existential risks. (You can find more detail here.)

Other causes

There are many other promising causes that, while not currently the primary focuses of the effective altruism community, are plausible candidates for having a big impact. These include:

- Improvements to the scientific establishment, such as greater transparency and replication of results
- Researching mental health and neurological disorders, particularly depression and anxiety, and improving access to treatment in low-income countries
- Tobacco control
- Prevention of road traffic injuries
- US criminal justice reform
- International migration and trade policy reform

Of course, it's likely that we have overlooked some very important causes. So one way to have a huge impact might be to find an opportunity to do good that's potentially high-impact, but that everyone else has missed. For this reason, global priorities research is another key cause area.

What does this mean for you?

Which career?

For most of us, a significant amount of our productive waking life — on average over 80,000 hours — is spent working. This is an enormous resource that can be used to make the world better. If you can increase your impact by just 1%, that's equivalent to 800 hours of extra work.

First, you need to consider which problem you should focus on. Some of the most promising problems appear to be safely developing artificial intelligence, improving biosecurity policy, or working toward the end of factory farming.

Next, you need to consider the most effective way for you to address the problem. At this point, it is useful to consider multiple approaches. Organizations working on these problems generally need more policy analysts, researchers, operations staff, and managers. But you should also consider more unconventional approaches, like being an assistant to another high-impact individual. Finally, you might want to consider whether you can have a bigger impact by doing community building, to persuade other people to address the problem. You can see some of the best opportunities right now on this job board. Senior hires are particularly in demand, but you can also often find graduate positions.

If you can't work directly on important problems now, consider building skills that allow you to do so in the future, or "earning to give" – taking a high-earning career and donating most of your salary.

But which role will be the best fit for you? When you donate to a charity, your personal attributes don't affect the value of your donation. But when it comes to choosing a career, your personal fit with a job is very important. This doesn't mean that you should just follow the standard career advice, and "follow your passion." Passion is much less important than you might think. But it is important to find a job that you excel at: you can best do this by trying out lots of options, and getting feedback about what you are good at.

80,000 Hours is an organization dedicated to helping people figure out in which careers they can do the most good. They provide a guide to the most important considerations relevant to career choice, and a set of tools to help motivated people make decisions. They have also conducted a large number of career reviews across a wide range of fields.

Which charity?

One of the easiest ways that a person can make a difference is by donating money to organizations that work on some of the most important causes. Monetary donations allow effective organizations to do more good things, and are much more flexible than time donations (like volunteering).

Most of us don't realize just how rich we are in relative terms. People earning professional salaries in high-income countries are normally in the top 5% of global incomes. This relative wealth presents an enormous opportunity to do good if used effectively.

Donating effectively

Some organizations associated with the effective altruism community seek out the most effective causes to donate to, backing up their recommendations with rigorous evidence.

One of the easiest ways to give to effective charities is through Effective Altruism Funds (note that this is a project that we run, so we might be biased in its favor!), EA Funds allows you to donate to one of four major cause areas. An expert in that cause area will then decide which organization could most effectively use the fund's donations. EA Funds also allows you to donate directly to a range of promising organizations.

You might also want to look at GiveWell's in-depth research into charities working within global health and development. You can see their current recommendations here.

Yet another place to search for promising donation targets is this list of organizations which have received a grant from the Open Philanthropy Project. The Open Philanthropy Project is a large grant-maker which uses effective altruist principles to find promising giving opportunities.

Pledging to give

It's easy to intend to give a significant amount to charity, but it can be hard to follow through. One way we can hold ourselves to account is to take a public pledge to give.

Giving What We Can has a pledge that asks people to give 10% or more of their lifetime income to the organizations that will make the biggest improvements to the world. The Life You Can Save has a similar pledge, starting at 1% of annual income, given to organizations fighting the effects of extreme poverty.

Get involved in the community

There's already a growing community of people who take these ideas seriously, and are putting them into action. Since 2009, more than 3,000 people have taken the Giving What We Can pledge to donate 10% of their lifetime income to the most effective organizations. Hundreds of people have made high-impact career plan changes on the basis of effective altruism. And there are over a hundred local effective altruism meet-up groups.

Taking action

If you're inspired by the ideas of effective altruism there are many ways you can take action:

- Read more
- Find a fulfilling career that does good
- Attend an effective altruism conference
- Find your local meet-up group

Efficient Charity — Do Unto Others

By Scott Alexander, 2013

This article was written in 2013, and estimates for the effectiveness of interventions have changed since then. We recommend visiting Givewell.org for more up to date estimates. This article was originally published as 'Efficient Charity: Do Unto Others' on lesswrong.com by Scott Alexander.

Imagine you are setting out on a dangerous expedition through the Arctic on a limited budget. The grizzled old prospector at the general store shakes his head sadly: you can't afford everything you need; you'll just have to purchase the bare essentials and hope you get lucky. But what is essential? Should you buy the warmest parka, if it means you can't afford a sleeping bag? Should you bring an extra week's food, just in case, even if it means going without a rifle? Or can you buy the rifle, leave the food, and hunt for your dinner?

And how about the field guide to Arctic flowers? You like flowers, and you'd hate to feel like you're failing to appreciate the harsh yet delicate environment around you. And a digital camera, of course - if you make it back alive, you'll have to put the Arctic expedition pics up on Facebook. And a hand-crafted scarf with authentic Inuit tribal patterns woven from organic fibres! Wicked!

...but of course buying any of those items would be insane. The problem is what economists call opportunity costs: buying one thing costs money that could be used to buy others. A hand-crafted designer scarf might have some value in the Arctic, but it would cost so much it would prevent you from buying much more important things. And when your life is on the line, things like impressing your friends and buying organic pale in comparison. You have one goal - staying alive - and your only problem is how to distribute your resources to keep your chances as high as possible. These sorts of economics concepts are natural enough when faced with a journey through the freezing tundra.

But they are decidedly not natural when facing a decision about charitable giving. Most donors say they want to "help people". If that's true, they should try to distribute their resources to help people as much as possible. Most people don't. In the "Buy A Brushstroke"campaign, eleven thousand British donors gave a total of 550,000 pounds to keep the famous painting "Blue Rigi" in a UK museum. If they had given that 550,000 pounds to buy better sanitation systems in African villages instead, the latest statistics suggest it would have saved the lives of about one thousand two hundred people from disease. Each individual $50 donation could have given a year of normal life back to a Third Worlder afflicted with a disabling condition like blindness or limb deformity..

Most of those 11,000 donors genuinely wanted to help people by preserving access to the original canvas of a beautiful painting. And most of those 11,000 donors, if you asked, would say that a thousand people's lives are more important than a beautiful painting, original or no. But these people didn't have the proper mental habits to realize that was the choice before them, and so a beautiful painting remains in a British museum and somewhere in the Third World a thousand people are dead.

If you are to "love your neighbor as yourself", then you should be as careful in maximizing the benefit to others when donating to charity as you would be in maximizing the benefit to yourself when choosing purchases for a polar trek. And if you wouldn't buy a pretty picture to hang on your sled in preference to a parka, you should consider not helping save a famous painting in preference to helping save a thousand lives.

Not all charitable choices are as simple as that one, but many charitable choices do have right answers. GiveWell.org, a site which collects and interprets data on the effectiveness of charities, predicts that antimalarial drugs save one child from malaria per $5,000 worth of medicine, but insecticide-treated bed nets save one child from malaria per $500 worth of netting. If you want to save children, donating bed nets instead of antimalarial drugs is the objectively right answer, the same way buying a $500 TV instead of an identical TV that costs $5,000 is the right answer. And since saving a child from diarrheal disease costs $5,000, donating to an organization fighting malaria instead of an organization fighting diarrhea is the right answer, unless you are donating based on some criteria other than whether you're helping children or not.

Say all of the best Arctic explorers agree that the three most important things for surviving in the Arctic are good boots, a good coat, and good food. Perhaps they have run highly unethical studies in which they release thousands of people into the Arctic with different combination of gear, and consistently find that only the ones with good boots, coats, and food survive. Then there is only one best answer to the question "What gear do I buy if I want to survive" - good boots, good food, and a good coat. Your preferences are irrelevant; you may choose to go with alternate gear, but only if you don't mind dying.

And likewise, there is only one best charity: the one that helps the most people the greatest amount per dollar. This is vague, and it is up to you to decide whether a charity that raises forty children's marks by one letter grade for $100 helps people more or less than one that prevents one fatal case of tuberculosis per $100 or one that saves twenty acres of rainforest per $100. But you cannot abdicate the decision, or you risk ending up like the 11,000 people who accidentally decided that a pretty picture was worth more than a thousand people's lives.

Deciding which charity is the best is hard. It may be straightforward to say that one form of antimalarial therapy is more effective than another. But how do both compare to financing medical research that might or might not develop a "magic bullet" cure for malaria? Or financing development of a new kind of supercomputer that might speed up all medical research? There is no easy answer, but the question has to be asked.

What about just comparing charities on overhead costs, the one easy-to-find statistic that's universally applicable across all organizations? This solution is simple, elegant, and wrong. High overhead costs are only one possible failure mode for a charity. Consider again the Arctic explorer, trying to decide between a $200 parka and a $200 digital camera. Perhaps a parka only cost $100 to make and the manufacturer takes $100 profit, but the camera cost $200 to make and the manufacturer is selling it at cost. This speaks in favor of the moral qualities of the camera manufacturer, but given the choice the explorer should still buy the parka. The camera does something useless very efficiently, the parka does something vital inefficiently. A parka sold at cost would be best, but in its absence the explorer shouldn't hesitate to choose the parka over the camera. The same applies to charity. An antimalarial net charity that saves one life per $500 with 50% overhead is better than an antidiarrheal drug charity that saves one life per

$5000 with 0% overhead: $10,000 donated to the high-overhead charity will save ten lives; $10,000 to the lower-overhead will only save two. Here the right answer is to donate to the antimalarial charity while encouraging it to find ways to lower its overhead. In any case, examining the financial practices of a charity is helpful but not enough to answer the "which is the best charity?" question.

Just as there is only one best charity, there is only one best way to donate to that charity. Whether you volunteer versus donate money versus raise awareness is your own choice, but that choice has consequences. If a high-powered lawyer who makes $1,000 an hour chooses to take an hour off to help clean up litter on the beach, he's wasted the opportunity to work overtime that day, make $1,000, donate to a charity that will hire a hundred poor people for $10/hour to clean up litter, and end up with a hundred times more litter removed. If he went to the beach because he wanted the sunlight and the fresh air and the warm feeling of personally contributing to something, that's fine. If he actually wanted to help people by beautifying the beach, he's chosen an objectively wrong way to go about it. And if he wanted to help people, period, he's chosen a very wrong way to go about it, since that $1,000 could save two people from malaria. Unless the litter he removed is really worth more than two people's lives to him, he's erring even according to his own value system.

...and the same is true if his philanthropy leads him to work full-time at a nonprofit instead of going to law school to become a lawyer who makes $1,000 / hour in the first place. Unless it's one HELL of a nonprofit.

The Roman historian Sallust said of Cato "He preferred to be good, rather than to seem so". The lawyer who quits a high-powered law firm to work at a nonprofit organization certainly seems like a good person. But if we define "good" as helping people, then the lawyer who stays at his law firm but donates the profit to charity is taking Cato's path of maximizing how much good he does, rather than how good he looks.

And this dichotomy between being and seeming good applies not only to looking good to others, but to ourselves. When we donate to charity, one incentive is the warm glow of a job well done. A lawyer who spends his day picking up litter will feel a sense of personal connection to his sacrifice and relive the memory of how nice he is every time he and his friends return to that beach. A lawyer who works overtime and donates the money online to starving orphans in Romania may never get that same warm glow. But the concern with a warm glow is, at root, concern about seeming good rather than being good - albeit seeming good to yourself rather than to others. There's nothing wrong with donating to charity as a form of entertainment if it's what you want - giving money to the Art Fund may well be a quicker way to give yourself a warm feeling than seeing a romantic comedy at the cinema - but charity given by people who genuinely want to be good and not just to feel that way requires more forethought.

It is important to be rational about charity for the same reason it is important to be rational about Arctic exploration: it requires the same awareness of opportunity costs and the same hard-headed commitment to investigating efficient use of resources, and it may well be a matter of life and death. Consider going to www.GiveWell.org and making use of the excellent resources on effective charity they have available.

Prospecting for Gold

By Owen Cotton-Barratt, 2016

In his 2016 talk, Oxford University's Owen Cotton-Barratt discusses how effective altruists can improve the world, using the metaphor of someone looking for gold. He discusses a series of key effective altruist concepts, such as heavy-tailed distributions, diminishing marginal returns, and comparative advantage. Watch the video by clicking on the image below. The below transcript is edited for readability.

Prospecting for Gold (Keynote) | Owen Cotton-Barratt | EAGxOxford 2016

Effective altruism as mining for gold

The central metaphor that is going to be running through my talk is 'effective altruism as mining for gold'. And I'm going to keep on coming back to this metaphor to illustrate different points. Gold, here, is

Figure 1: Mining for gold

standing in for whatever it is that we truly value. Some things that we might value include making more people happy and well-educated. Or trying to avert a lot of suffering. Or trying to increase the probability that humanity makes it out to the stars. When you see gold, take a moment to think about what you value. Many people won't just value one particular thing. However, do think about what you care about and put that in place of the gold. This way, there are lots of observations we can make.

This [figure 2] is a photo of Viktor Zhdanov, and I learned about him by reading Will MacAskill's book 'Doing Good Better'. He was a Ukrainian biologist, who was instrumentally extremely important in getting an eradication program for smallpox to occur. As a result, he probably was counterfactually responsible for saving tens of millions of lives.

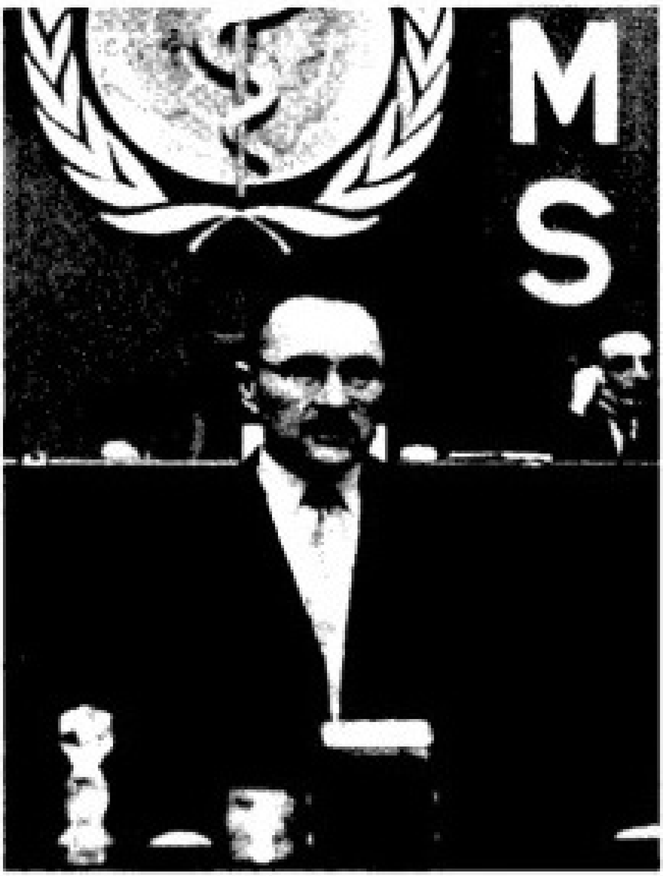


Figure 2: Victor Zhdanov

Obviously, we don't all achieve this. But by looking at examples like this, we can notice that some people manage to get a lot more gold, manage to accomplish a lot more of whatever we altruistically value, than others. And that is reason enough to make us ask questions like: what is it that gives some people better opportunities than others? How can we go and find opportunities like that?

Techniques for finding gold

Figure 3: Techniques for finding gold

Elsewhere in this conference, there are going to be treasure maps and discussions of where the gold is. I'm not going to do that in this talk. I'm instead going to be focusing on the tools and techniques that we can use for locating gold, rather than trying to give my view of where it is directly.

I want to say a little bit about why I'm even using a metaphor –because we care about these things. We care about a lot of these big, complicated, valuable things. Why would I try and reduce that down to gold? Well, it is because of where I want the focus of this talk to be. I want the focus to be on techniques, tools and approaches that we can use. And if you have complex values, these would just keep on pulling your attention. But a lot of the things that we might do to try to identify where valuable things are, and how to go and achieve them, are constant, regardless of what the valuable thing is. So, by replacing them with a super simple stand-in for value, I think it helps to put the focus on this abstract layer that we are putting on top of that.

Gold is unevenly spread

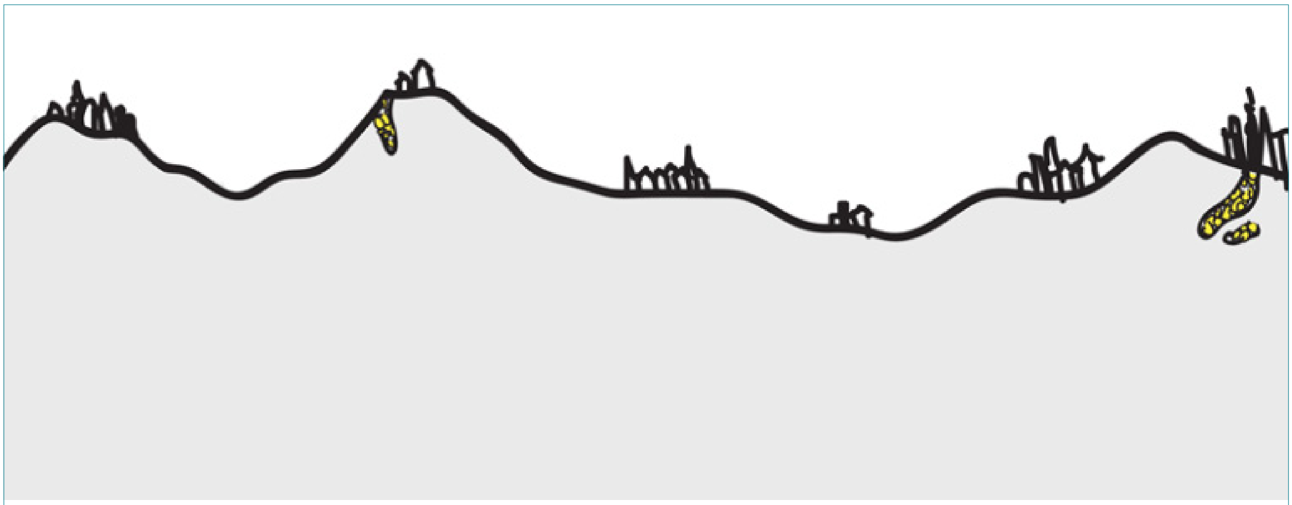


Figure 4: Gold is unevenly spread

The first thing I'm going to talk about is the fact that gold, like literal gold, is pretty unevenly spread around the world. There are loads of places with almost no gold at all, and then there are a few places where there's a big seam of gold running into the ground. This has some implications. One is that we would really like to find those seams.

Heavy-tailed distributions

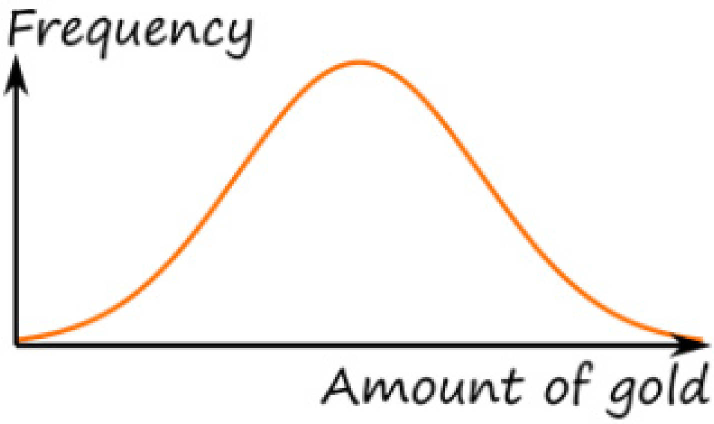


Figure 5: Non-heavy tailed distributions

Another is about sampling. For some quantities, say if I want to know roughly how tall people are, sampling five people, measuring their height and saying, "Well, the average is probably like that" is not a bad methodology. However, if I want to know on average how much gold there is in the world, sampling five random places, and measuring that, is not a great methodology, because it's quite likely that I'll find five places where there's no gold, and I'll significantly underestimate. Or possibly, one of them will have a load of gold, and now I'll have a massively inflated sense of how much gold there is in the world.

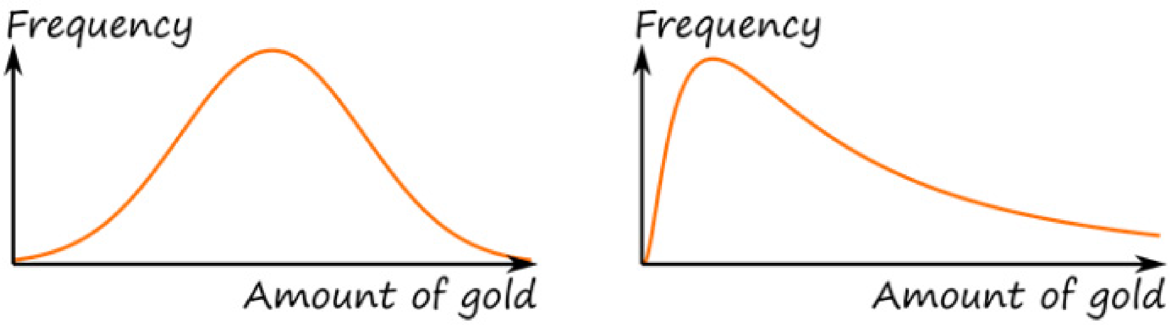


Figure 6: Non-heavy tailed distribution vs. heavy tailed distribution

This is a statistical property that loosely gets called having a heavy tail on the distribution. This here on the left is a distribution without a heavy-tail [Figure 6]. There is a range of different amounts of gold in different places, but none of them has massively more or massively less than typical.

On the right, in contrast, is a heavy-tail distribution. It looks similar-ish to the one on the left-hand side, but there's this long tail, getting up to vast amounts of gold, where the probabilities aren't dying off very fast. This has implications.

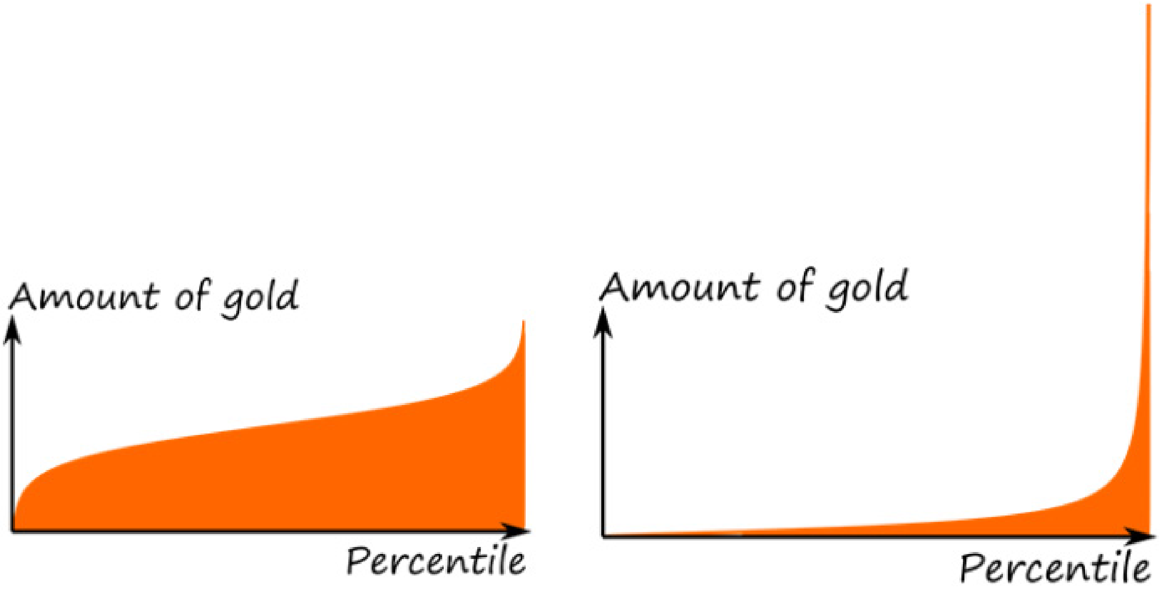


Figure 7: Heavy-tailed distributions

Here is another way of looking at these distributions [Figure 7]. In this case, I've arranged, going from left to right, the places in order of increasing amounts of how much gold they have. The percentiles are on the horizontal axis and the amount of gold on the vertical axis. In this case, the area beneath the graph is coloured in. That is because that quantity - that area - is meaningful. It corresponds to the total amount of where that gold is. So, on the left of the distribution that is not heavy-tailed, I can see that the gold is fairly evenly spread across a lot of different places. If we want to just get most of the gold, what is important is getting to as many different places as possible.

Solar power is like this. Sure, some places get more sunlight than other places, but the amount of solar power you generate depends more on how many total solar panels you have, than on exactly where you place them.

Over on the right, though, we have a distribution where you can see a lot of the area that spikes on the right-hand side. This just means that a lot of the gold, and if this is an actual stand-in for something that we value, a lot of what is valuable comes in this extreme of the distribution of things, which are just unusually good.

Literal gold, I think, is distributed like this. Disclaimer, I'm not a geologist, I don't know anything about gold, but I understand that this is right. We might ask, is this also true of opportunities to do good in the world? Here is some support for this.

Heavy-tailed property in the wild

Arise naturally

- e.g. log-normal, power law distributions

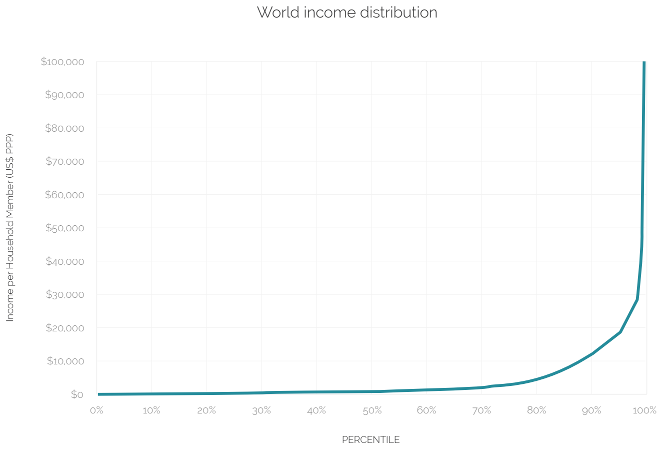


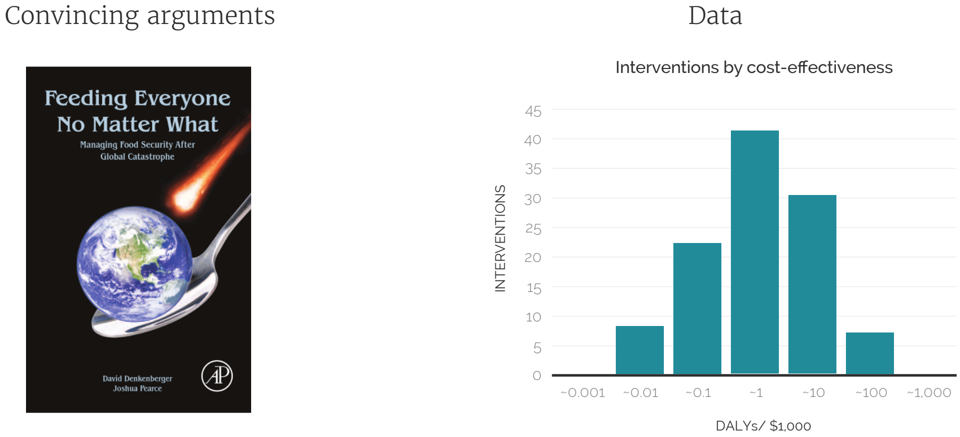
Figure 8: Heavy-tailed property in the wild

When we look into the world, and it's pretty complex, we do see distributions with this heavy-tail property coming up in a lot of different places [figure 8]. There are some theoretical reasons to expect certain types of distributions to arise. And this is also the case empirically if we go and look at something like income distributions around the world. Again this is a percentile version, and you can see the spike in the graph.

So this is what we find if we just go and look at the world. Obviously, there are lots of things as well that don't have this property. But the more we look at things that are complex systems with lots of interactions, the degree to which we see this property increases. That is a big feature of lots of ways that we try and interact to improve the world.

I can also just try and look explicitly at opportunities to do good. And I can see a couple of reasons why I personally am convinced that we get some of this property. So, one reason is just convincing arguments. If I care about stopping people starving, and I do care about stopping people starving, I could ask: should I be interested in direct famine relief and trying to get food to people who are starving today? I can compare this to something more speculative. I personally have been convinced by the arguments in this book, that it would be more effective to focus on researching solutions for feeding vast numbers of people if agriculture collapses. It's pretty extreme. It's not something we usually think about, but I think that the argument basically checks out. I'm just limiting myself here to trying to feed people, and one of the mechanisms looks much more effective than the other.

Heavy-tailed property in opportunities for good



Convincing arguments

Data

Figure 9: Heavy-tailed property in opportunities for good

I can also look at data [Figure 9]. This is data from DCP2, which has tried to estimate the cost-effectiveness of lots of different developing world health interventions. The x-axis is on a log scale, so these have been put into buckets, and each column is on average ten times more effective to than the one on its left. Here the rightmost column is about 10,000 times more effective than the leftmost column. And this was just within one area where we have managed to get good enough data that we can go and estimate these things. There is just a very wide range of cost effectivenesses.

The implications of this are that if we want to go and get gold, we really should focus on finding seams. In some cases, it might give us a surprising conclusion, say you discover that something is at the 90th percentile, that might make us less excited about it. Because before we knew anything, it might have been anywhere on the distribution. And if most of the possible value of it comes from it being up at the 99th percentile, then discovering it's only at the 90th percentile could actually be a bad thing. I mean, it's a good thing to discover, but it makes us think less well of it. Now, that's if you've got a fairly extreme distribution, but it's interesting to see how you can get these kinds of counterintuitive properties.

Another implication is that a kind of naïve empiricism, "we'll just do a load of stuff and see what comes out best", isn't going to be enough for us in judging this, because of this sampling issue. We can't go and sample enough times and measure the outcomes well enough to judge how effective it is really going to be.

To maximize gold…

Figure 10: Maximising gold

If we want to get as much gold as possible, we want to go to a place where there is lots of gold. We want to have the right tools for getting the gold out, and we want to have a great team that is going to be using those tools. I think that we can port this analogy over to opportunities to doing good as well. We can roughly measure the effectiveness of the area or type of thing that we're doing, and the effectiveness of the intervention that we're implementing to create value in that area, relative to other interventions in the area. We can measure the effectiveness of the team or the organisation which is running the implementation, relative to how well other teams might implement such an intervention.

Value is roughly multiplicative

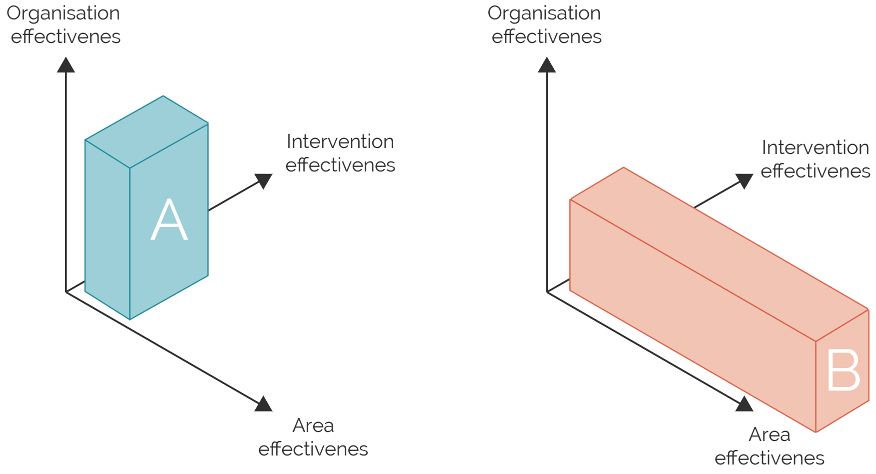

Organisation effectivenes
Intervention effectivenes
A
Area effectivenes
Organisation effectivenes
Intervention effectivenes
B
Area effectivenes

Figure 11: Value is roughly multiplicative

And if you have these things then the total value that you are going to be getting is equal to the product of these. I've represented it here by volume, and we want to be maximizing the volume [Figure 11]. That means we're going to want to be trying to do reasonably well on each of the different dimensions. At least, not terribly on any of the dimensions. Some implications here might be that if we have an area and an intervention that we're excited about, but we can only find a kind of mediocre team working on it, it may be better not just to support them, but to try and get somebody else working on it. Or to do something to really improve that team. Similarly, we might not want to support even a great team, if they're working in an area that doesn't seem important.

Recognising gold

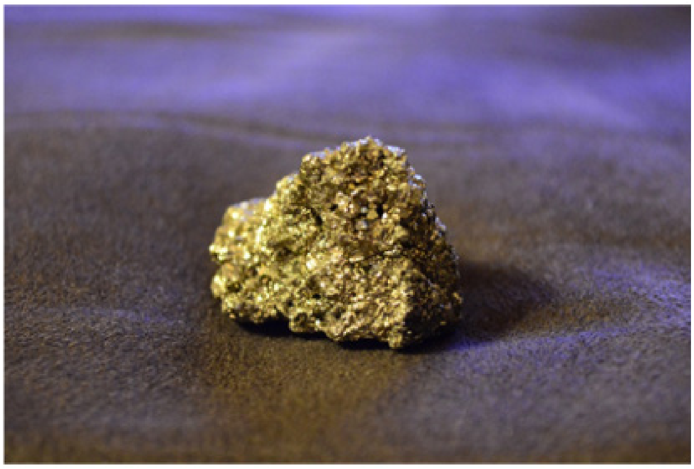


Figure 12: Gold

In the next section, I am going to talk about the tools and techniques for identifying where in the world gold is. A nice property of literal gold is that when you dig it up, you're pretty sure that you can recognise "yes I have gold". We often have to deal with cases where we don't have this. We don't have the gold, so we have to carefully try to infer its existence, by using different tools. This fact is like the dark matter of value.

This fact increases the importance of actually applying those tool diligently. Actually, the picture I showed you [Figure 12] was iron pyrite, not gold. So just because somebody says, "Hey, this is gold," doesn't mean we should always take people's word for it. It does provide some evidence, but we have motivation for wanting to have great tools for identifying particularly valuable opportunities. And being able to

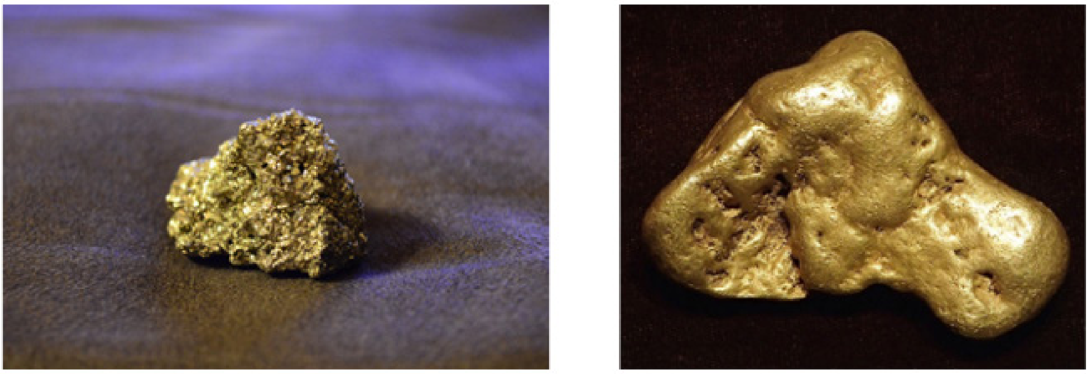


Figure 13: Not gold (left) versus gold (right)

differentiate and say, “Okay, actually this thing, although it has some aspects of value, may not be what we want to pursue.”

Running out of easy gold

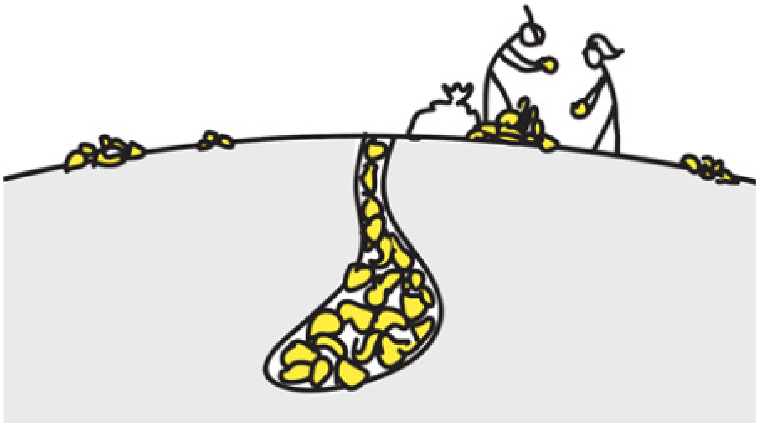


Figure 14: Running out of easy gold

If you first go to an area where nobody has been before, then the seams of gold that are running through the ground have often been eroded a little bit, and you can have little nuggets of gold just lying around on the ground, and it’s extremely easy to get gold. So you have some people go in, they do this for a bit, and they run out of all the gold on the ground.

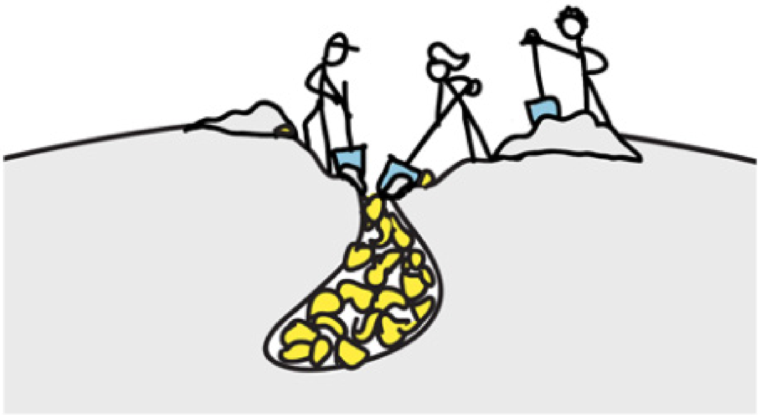


Figure 15: Running out of easy gold (2)

And now, if they want to get more gold, maybe more people come along, they bring some shovels, and it is a bit more work, but you can still get gold out [figure 15].

And then you dig deep enough, and you can’t just get in with shovels anymore, so you need bigger teams and heavier machinery to get gold out [figure 16]. You can still get gold, but it’s more work for each little bit, for each nugget that you’re getting out. This is the general phenomenon of ‘diminishing returns’ on work that you’re putting in. This concept comes up in a lot of different places, and so it is worth having an understanding of it.

By the way, this, like several of the things I’m going to be talking about, is a concept, which is native to

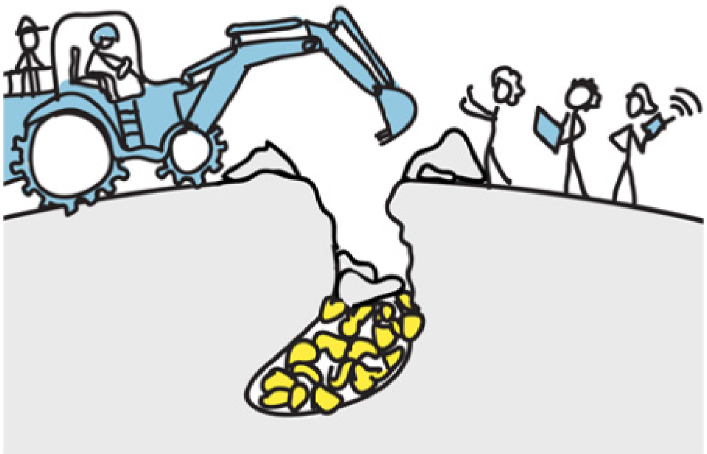


Figure 16: Running out of easy gold (3)

economics. And in some cases, I'm merely just pulling this from economics, and in some cases, there's a little bit more modification on the concept.

For instance, I think that we get this in global health. I understand that 15 or 20 years ago, mass vaccinations were extremely cost-effective and probably the best thing to be doing. Then the Gates Foundation has come in and funded a lot of the mass vaccination interventions. Now, the most cost-effective intervention is less cost-effective than mass vaccinations. That is great because we have taken those low hanging fruit. Or similarly, if in AI safety, writing the first book on superintelligence is a pretty big deal. Writing the 101st book on superintelligence is just not going to matter as much.

So, a minute ago, I talked about how we could factor the effectiveness of organisations into the area which it was working on, the intervention it is pursuing, and the team working on it. Now, I'm going to focus on that first one, trying to assess the area. And I'm going to give a further factorisation, splitting this into three different things.

Scale

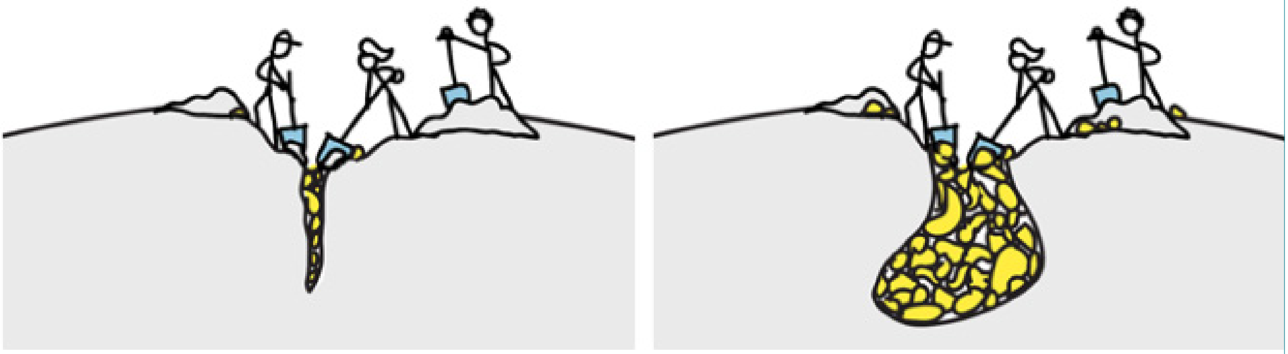


Figure 17: Scale

The first of these dimensions is scale. All else being equal, we would prefer to go somewhere where there is a lot of gold, rather than a little bit of gold. And probably per unit efforts, we are going to get more gold, if we do that.

Tractability

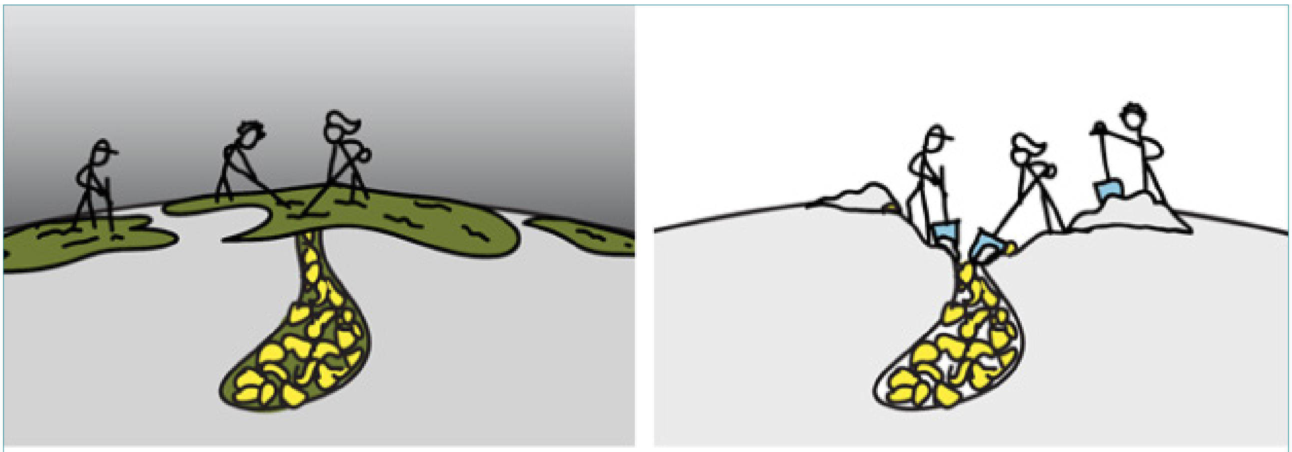


Figure 18: Tractability

Second: tractability. We'd like to go somewhere where we make more progress per unit work. Somewhere, where it's nice and easy to dig the ground, rather than trying to get your gold out of a swamp.

Uncrowdedness

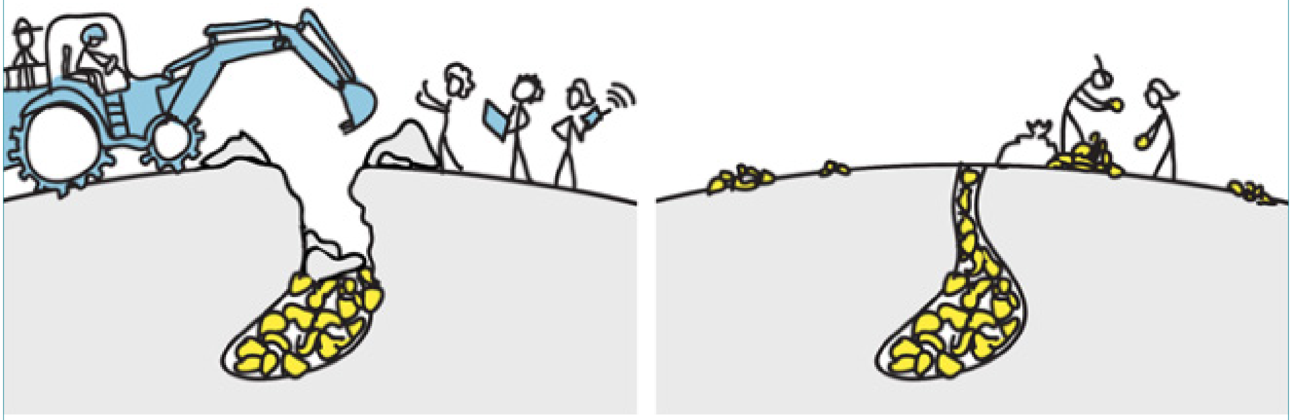


Figure 19: Uncrowdedness

And third is uncrowdedness. This has sometimes been called neglectedness. I think that term is a bit confusing. It's a bit ambiguous because sometimes people use 'neglectedness' to mean that this is an area which we should allocate more resources to. What I mean here is that there aren't many people looking at it. All else being equal, we would rather go to an area where people haven't already picked up the nuggets of gold on the ground, than one where they have. And now the only gold remaining is quite hard to extract.

Ideally, of course, we'd like to be in the world where there is loads of gold, that's easier to get out, and nobody has taken any of it. But, we are rarely going to be in that exactly ideal circumstance. So one question is: how can we trade these off against each other? I'm going to present one attempted way to try and make that precise. I've allowed myself one equation in this talk. This is it [figure 20].

Scale, Tractability, Uncrowdedness

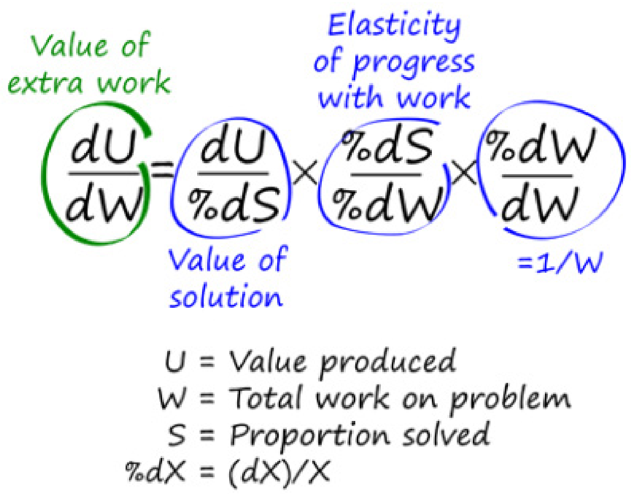


Figure 20: Scale, tractability, uncrowdedness

If you're not used to thinking in terms of derivatives, just ignore the 'ds' here [Figure 20]. But on the left is the value of a little bit of extra work, so this is what we care about if we're trying to assess which of these different areas we should do more work on.

On the right is a factorization. This is mathematically trivial, I've just taken this one expression, and I've added a load more garbage. And on the face of it, it looks like I've made things a lot worse. I can only justify this, if it turns out that these terms I've added in, which cancel each other out, actually mean that the right-hand side here is easier to interpret or easier to measure. So I'm going to present a little bit of a case for why I think it is.

The first term is measuring the amount of value you get for, say, solving an extra one percent of a solution. And that roughly tracks how much of a big deal the whole problem that you're looking at is, the whole area. I think that is a pretty precise version of the notion of scale.

The second one is a little bit more complicated. It is an elasticity here, which is a technical term. It's a pretty useful and general term (go look it up on Wikipedia, if you're interested). Here, it is measuring, for a proportional increase in the amount of work that's being done, what proportion of a solution that gives you.

The final term just cancels to one over the total amount of work being done. So that is very naturally a measure of uncrowdedness.

People have talked about this kind of scale, tractability, uncrowdedness framework for a few years without having a precise version. That means that people have given different characterizations of the different terms, and I think there have been a few different versions of tractability, not all of them lining up with this exactly. But I think that this idea of it - measuring how much more work gets you towards a solution - is fairly well captured here.

All three dimensions matter

Figure 21: All three dimensions

I think that all of these dimensions matter. And again, that means we probably don't want to work on something, which does absolutely terribly on any of the dimensions. I'm not going to spend an hour helping a bee, even if nobody else is helping it and it would be pretty easy to help, because just the scale of it is pretty small. I don't think we should work on perpetual motion machines, even though basically nobody is working on it and it would be really fantastic if we succeeded. Because it seems like it's not tractable.

And this […] might give us a warning against actually working on climate change. Because at a global scale, that gets a lot of attention, as a problem.

I'm going to add some more caveats to that one. One is that this is going to be true while we think that there are other problems, which are just significantly more under-resourced. And another is that you might think that you have an exception if you have a much better way of making progress on the problem of climate change than typical work that is done on it.

Even so, I think maybe we should think it's a bit surprising that I'm making a statement like "climate change is not a high priority area." This just sounds controversial and we should be sceptical of this. But, I think that the term high priority is a little bit overloaded. And so I want to distinguish that a little bit.

Absolute and marginal priority

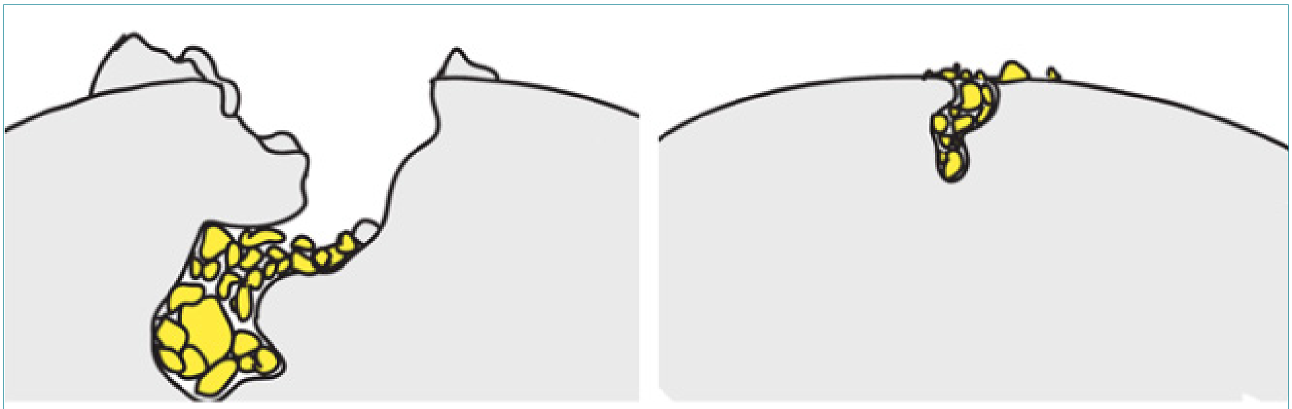


Figure 22: Absolute and marginal priority

If we have these two places where there's gold in the ground, and we say: "Where should we send people if we want to get gold?" The answer is going to depend. Maybe we send the first person to this place on the right, where there's only a little bit of gold, but it's really easy to get out. Then we send the next ten people to the place on the left, just because there's more total gold there. The first person will already have gotten most of the gold on the right. And we want more people total working on this place on the left. Which of these is the higher priority? Well, that just depends on which question you're asking.

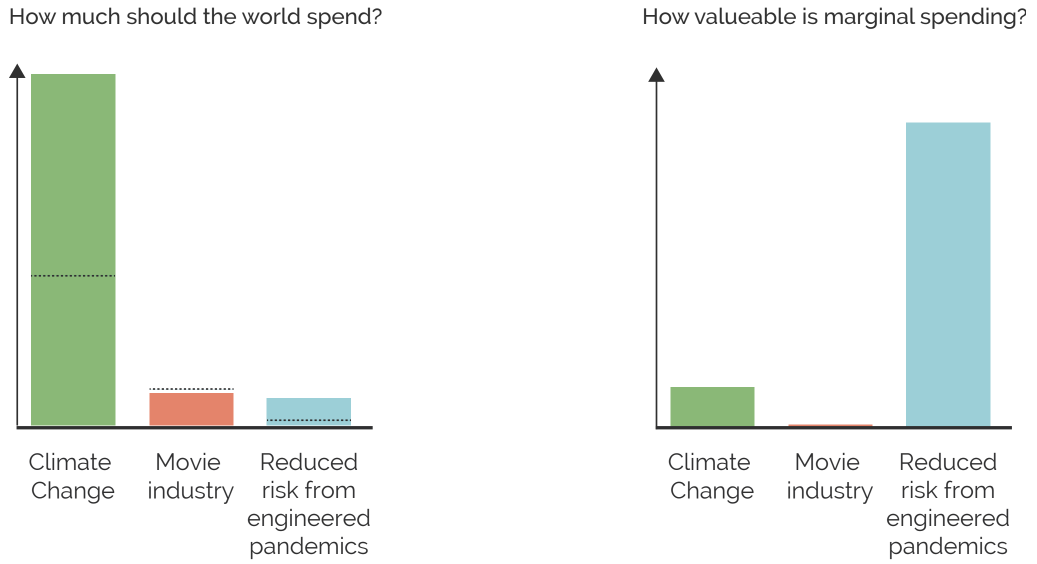


Figure 23: Absolute and marginal priority

These numbers are just made up off the top of my head, but we might have some distribution like this on the left. Here, we ask the question "How much should the world spend on this area, total?" and we get one distribution, where maybe climate change looks very big.

And if we instead ask, how valuable is marginal spending? The graph might look quite different because here it is significantly about how much is already being spent. You'll see some dotted black lines on the diagram on the left [Figure 23] – they might represent how much is already being spent. Then, the graph on the right is a function of all sorts of things, like how much should be spent in total, how much is already being spent and of course, what the marginal returns are – what the curve looks like there.

But I think that both of these are important notions, and which one we use should depend on what we are talking about. If we are having a conversation about what we as individuals or as small groups should do, I think it's appropriate to use this notion of marginal priority, of how much do extra resources help. If we're talking about what we collectively as a society or the world should do, I think it's often correct to talk about this kind of notion of absolute priority and how much resources ought to be invested in it, total.

Okay, for most of the things here, I've been extremely agnostic about what our view of value is. Just for this point, I'm going to start making more assumptions. I think quite a few people have the view that what we want to do is try and make as much value over the long term, as we can. Some people don't have that view, some people haven't thought about it. If you don't have that view, you can just treat this as a hypothetical: "Now I can understand what people with that view would think." If you haven't thought about it, go away and think about it, some time. It's a pretty interesting question, and I think it's an important question, and is worth spending some time on.

But, if we do care about creating as much value in the long term as possible, in our gold metaphor, that might mean wanting to get as much gold out of the ground eventually, as possible, rather than just trying to get as much gold out of that ground this year.

Long-term gold

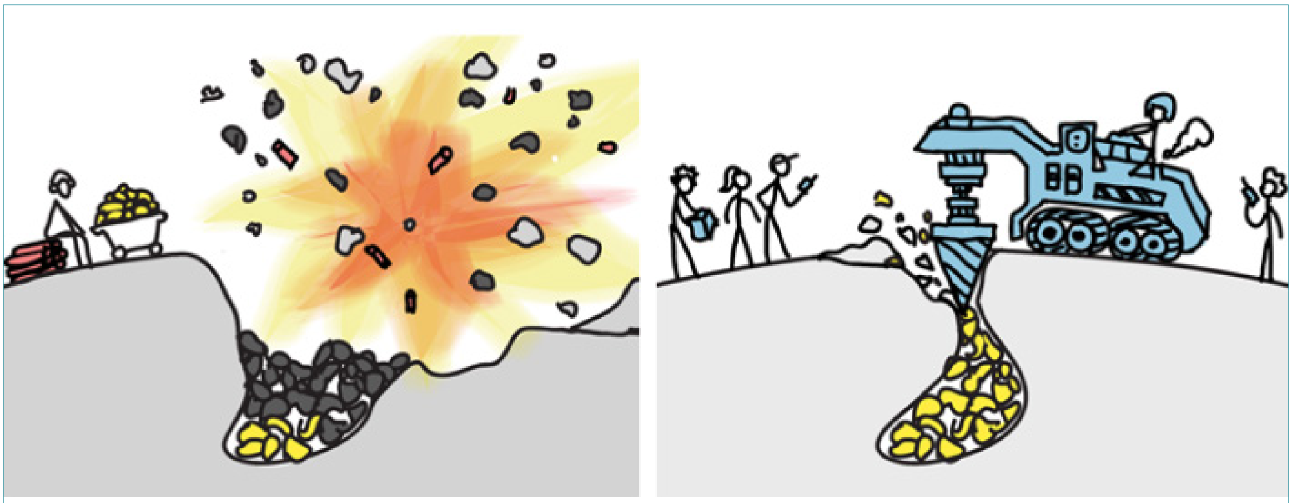


Figure 24: Long-term gold

And maybe we have some technologies which are destructive. So we can use dynamite and dynamite gets us loads of gold now, but it also blows up some gold, and now we never get that gold later. That could be pretty good if you are focusing just on trying to get gold in the short term. But it could be bad from this eventual gold perspective.

If we have different technologies that we can develop, maybe we can develop some that are also efficient but less destructive. And there are going to be some people in the world who do care about creating as much gold as possible in the short term. They are going to use whichever technology is the most efficient for that. And so one of the major drivers of how much gold is eventually extracted is the order in which the technologies are developed, and the sequencing. If we discover the dynamite first, people are going to go and have fun with their dynamite and they're going to destroy a lot of the gold. If we discover the drill first, then by the time dynamite comes along, people will go "Well, why would we use that? We have this fantastic drill."

Philosophers like Nick Bostrom have used this to argue for trying to develop societal wisdom and good institutions for decision making, before developing technologies or progress which might threaten the long-run trajectory of civilization. And also for trying to focus on differentially aiming to develop technologies which enhance the safety of new developments, rather than or before anything that's driving risk.

Working together

Now, I'm going to talk about how this is a collaborative endeavour. We're not just all, each of us individually, saying: "I need to work out where the most gold is. And that's most neglected, most tractable. I personally am just going to go and do that." Because there is a whole lot of people who are thinking like this, and there are more every year. I am really excited about this. I'm excited to have so many people here and also this idea that maybe in two years time, we'll have a lot more again.

But we need to work out how to cooperate. Largely, we have the same view or pretty similar views on what to value. Maybe some people think that silver matters, too – it's not just gold – but we all agree that gold matters. We're basically cooperating, here. We want to be able to coordinate and make sure that we're getting people working on the things which make the most sense for them.

Comparative advantage

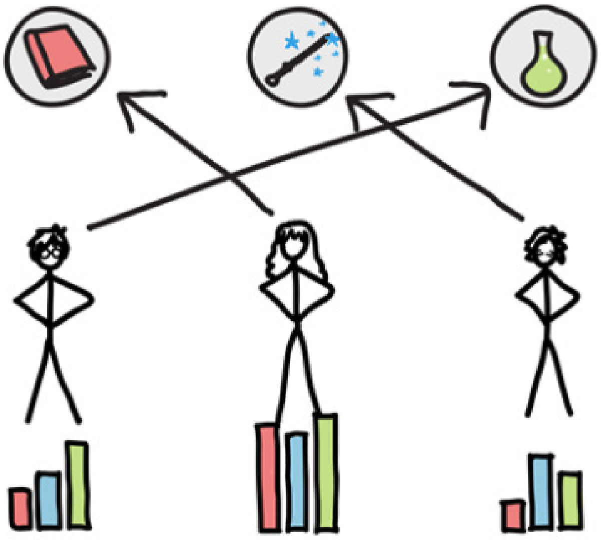


Figure 25: Comparative advantage

There is this idea of comparative advantage. On this graph [Figure 25], I have Harry, Hermione and Ron, and they have three tasks that they need to do to get some gold. They need to do some research, they need to mix some potions, and they need to do some wand work. Hermione is the best at everything, but she doesn't have a time turner, so she can't do everything. So we need to have some way of distributing the work. This is the idea of comparative advantage. Hermione has an absolute advantage on all of these tasks, but it would be a waste for her to go and work on the potions because Harry is not so bad at potions. And really, nobody else is at all good at doing the research in the library. So we should probably put her on this.

And this is a tool that we can use to help guide our thinking about what we should do as individuals. If I think that some technical domain and technical work is the most valuable thing to be doing, but I would be pretty mediocre at that, and I'm a great communicator? Then maybe I should go into trying to help technical researchers in that domain communicate their work to get more people engaged with it and bring in more fantastic people.

Comparative advantage at different levels

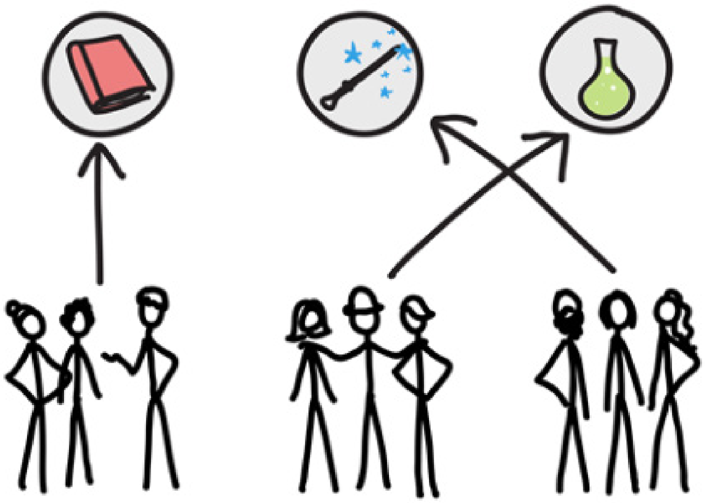


Figure 26: Comparative advantage at different levels

Now we have applied this at an individual level. We can also apply this at the group level. We can notice that different organizations or groups may be better placed to take different opportunities.

This is a bit more speculative, but I think we can also apply this at the time level. We can ask ourselves, "What are we, today, the people of 2016, particularly well suited to do, versus people in the past and people in the future?" We can't change what people in the past did. But we can make a comparison of what our comparative advantage is relative to people in the future. And if there is a challenge, if there were going to be some different possible challenges in the future that we need to meet, it makes sense that we should be working on the early ones. Because if challenges are coming in 2020, the people in 2025 just do not have a chance to work on that.

Another thing which might come here is that we have a position, perhaps, to influence how many future people there will be who are interested in and working on these challenges. We have more influence over that than people in that future scenario do, so should we think about whether that makes sense as a thing for us to focus on?

Building a map together

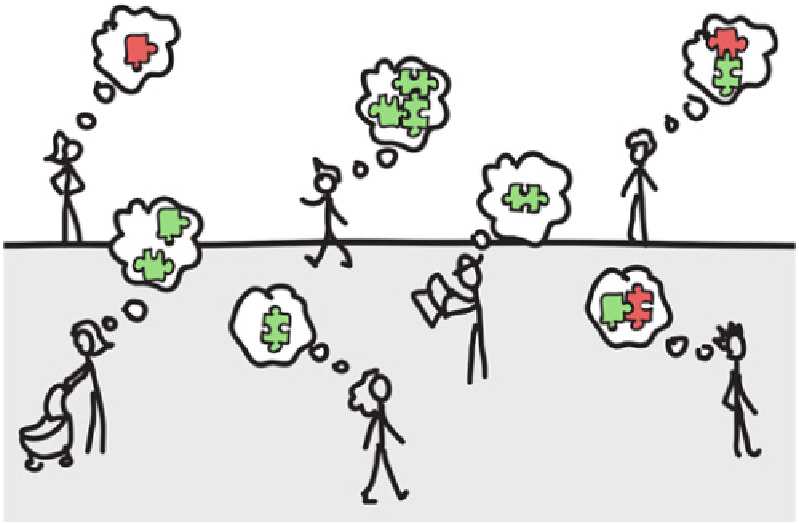


Figure 27: Building a map together

Another particularly important question is how to work stuff out. The world is big and complicated and messy. And we can't expect all of us, individually, to work out perfect models of it. In fact, it is too complicated for us to expect anybody to do this. So, maybe we're all walking around with the little ideas which, in my metaphor here, are puzzle pieces for a map to where the gold is. We want institutions for assembling these into a map. It's a bit complicated, because some people have puzzle pieces which are from the wrong puzzle, and they don't track where the gold is. Ideally, we would like our institutions to filter these out and only assemble the correct pieces to guide us where we want to go.

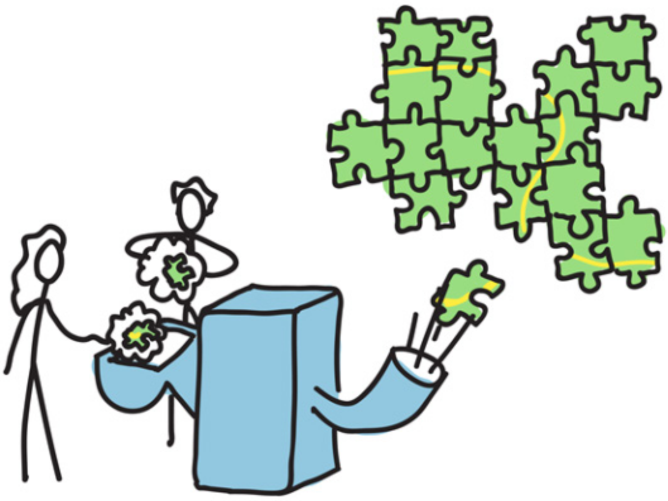


Figure 28: Building a map together (2)

As a society, we have had to deal with this problem in a number of different domains, and we have developed different institutions for doing this. There is the peer review process in science. Wikipedia does quite a lot of work aggregating knowledge. Amazon reviews aggregate knowledge that individuals have about which products are good. Democracy lets us aggregate preferences over many different people to try and choose what's going to be good.

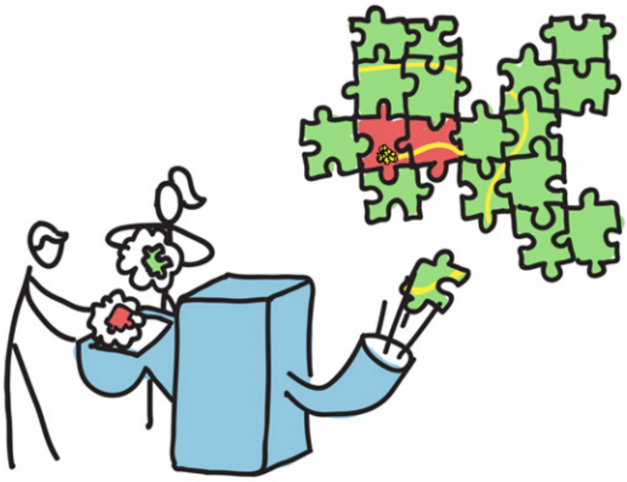


Figure 29: Building a map together (3)

Of course, none of these institutions is perfect. And this is a challenge. This is like one of those wrong puzzle pieces, which made it into the dialogue. And this comes up in the other cases as well. The replication crisis in parts of psychology has been making headlines recently. Wikipedia, we all know, sometimes gets vandalized, and you just read something which is nonsense. Amazon reviews have problems with people giving fake reviews, to make their product look good or other people's products look bad.

So, maybe it is the case that we can adapt one of these existing institutions for our purpose, which is trying to aggregate knowledge about what are the ways to go and do the most good. But maybe we want something a bit different, and maybe somebody in this room is going to do some work on coming up with valuable institutions for this. I actually think this is a really important problem. And it's one that is going to become more important for us to deal with as a community, as the community grows.

Good local norms

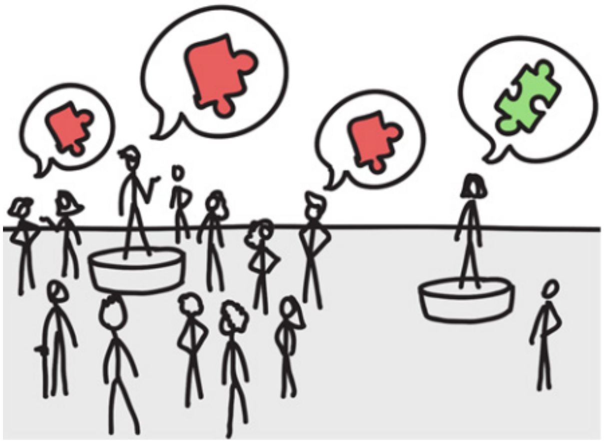


Figure 30: Good local norms

That was all about what are our global institutions for pulling this information together and aggregating it. Another thing, which can help us to move towards getting a better picture, is trying to have good local norms. So, we tell people the ideas that we have, and then other people maybe start listening. And sometimes it might just be that they listen based on the charisma of the person who is talking, more than based on the truthiness of the puzzle piece. But we would like to have ways of promoting the spread of good ideas, inhibiting the spread of bad ideas, and also encouraging original contributions. One way of trying to promote the spread of good ideas and inhibit bad ideas is just to rely on authority. We'll say, "Well, we've worked out this stuff. We're totally confident about this. And now we just won't accept anything else." But that isn't going to let us get new stuff.

Pay attention to why we believe things

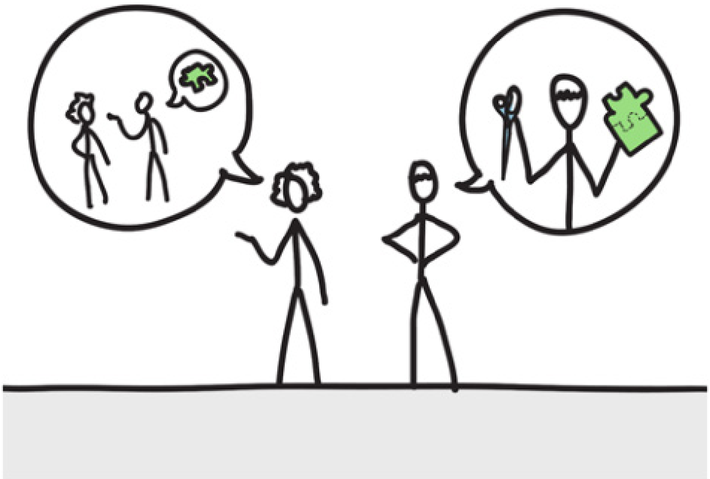


Figure 31: Pay attention to why we believe things

I think something to do here is to pay attention to why you believe something. Do you believe it because somebody else told you? Do you believe it because you have really thought this through carefully and worked it out for yourself? There is a blur between those. Often somebody tells you, and they give you some reasons. And you are thinking: "Oh, those reasons kind of check out," but you haven't deeply examined the argument yourself.

I think it's useful, to be honest with yourself about that. And then also to communicate it to other people. To let them know why it is. Is it the case that you believe this because Joe Bloggs has told you? And actually, Joe is a pretty careful guy, and he is pretty diligent about checking out his stuff, so you think it probably makes sense. You can just communicate that. Or is it that you cut out this puzzle piece yourself?

Now, cutting it out yourself doesn't necessarily mean we should have higher credence in it. I have definitely worked things out, and I have thought I have proved things before, and there was a mistake in my proof. So you can separately keep track of the level of credence you have in a thing, and why you believe it.

Also, our individual and collective reasons for believing things can differ. Here is this statement, that it costs about $3,500 to save a life from malaria. I think this is broadly believed across the effective altruism community. I think that collectively, the reason we believe this is that there have been a number of randomized control trials. And then some pretty smart, reasonable analysts at GiveWell have looked carefully at this, and they have dived into all the counterfactuals, and they have produced their analysis, and come to the conclusion: "On net, it looks like it's about $3,500."

Shortening the chain

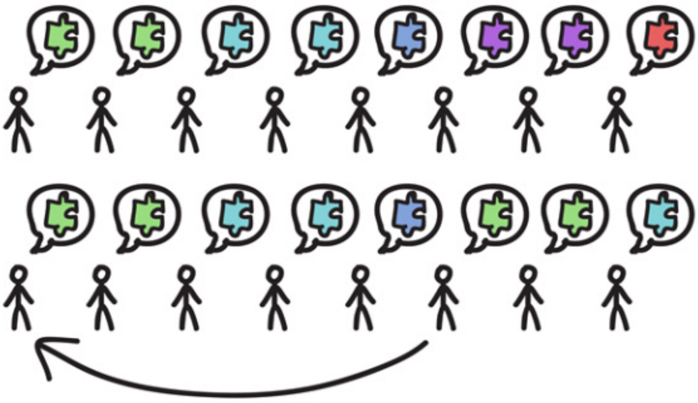


Figure 32: Shortening the chain

But that isn't why I believe it. I believe it because people have told me that the GiveWell people have done this analysis and they say it's $3,500. And they say, "Oh, yeah. I read it on the website." That was why I believed it until I started prepping for this talk when I went and read it on the website. Because I think that this is a bit more work for me, but it's doing a bit of value for the community. I'm shortening the chain of Chinese whispers, of passing this message along. And as things get passed along, it's possible that mistakes enter or just something isn't well grounded, and then it gets repeated. By going back and checking earlier sources in the chain, we can try to reduce that, and try to make ourselves more robustly confident in these statements.

Disagreement is an opportunity to learn

Figure 33: Disagreement is an opportunity to learn

Another thing that comes up is when you notice that you disagree with somebody.: If you're sitting down and talking with someone and they're saying something, and you think: "Well, that's obviously false." You can see perhaps that parts of their jigsaw puzzle are wrong. You could just dismiss what they have to say. But I think that that's often not the most productive thing to do. Because even if part of what they have to say is wrong, maybe they have some other part that is going into their thinking process which would fill a gap in your perspective on it, and help you to have a better picture of what's going on.

I often do this when I find that someone has a perspective that I think is unlikely to be correct. I'm interested in this process of how they get there and how they think about it. Partly this is just that people are fascinating and the way that people think is fascinating, so this is interesting. But I also think that it is polite and I think it is useful. I think it does help me to build a deeper picture of all the different bits of evidence that we collectively have.

Retrospective – What I believe and why

In this section, I'm going to put the stuff I've just been talking about into action. I've told you about a whole load of different things through this talk. But I didn't tell you much about exactly what my level of competence in these is, or why I believe these. So, I'm going to do that here.

I'm aware that nobody ever goes away from a talk saying, "Oh, that was so inspiring. The way she carefully hedged all her statements." But I think it's important. I would like people to go away from talks saying that. So I'm just going to do it.

Heavy-tailed distributions

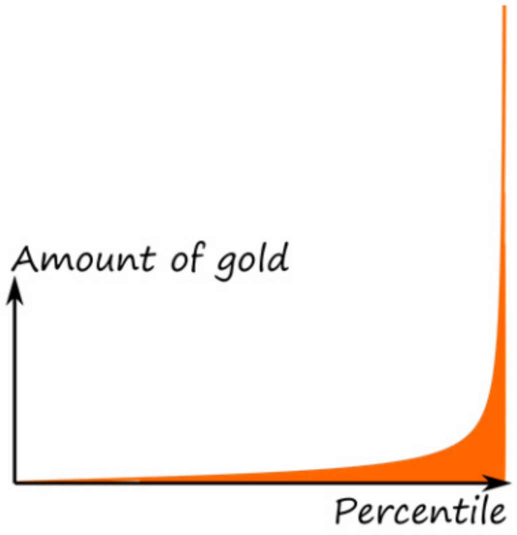


Figure 34: Heavy-tailed distributions

Heavy-tailed distributions: I think it's actually pretty robust that the kind of baseline distribution of opportunities in the world does follow something like this, a distribution with this heavy-tailed property. I think that just seeing this in many different domains and understanding some of the theory behind why it should arise makes it extremely likely. I think that there's an open empirical question to exactly how far that tail goes out. Heavy-tailedness isn't just a binary property; it's a continuum. Anders Sandberg is going to be talking more about this, I think, later today.

Digression: Altruistic market efficiency

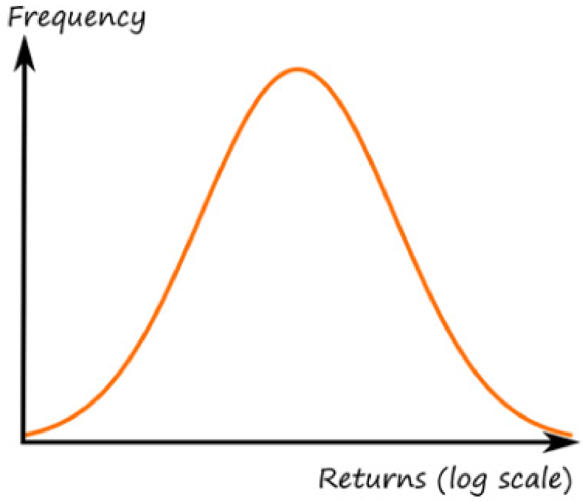


Figure 35: Altruistic market efficiency

But, there is an important caveat here. This is the only one of these I've allowed myself, a digression. It is that there is a mechanism which might push against that, which is people seeking out and taking the best opportunities. If people are pretty good at identifying the best opportunities, and they are uniformly seeking out and taking them, then the best things that are left might not be so much better.

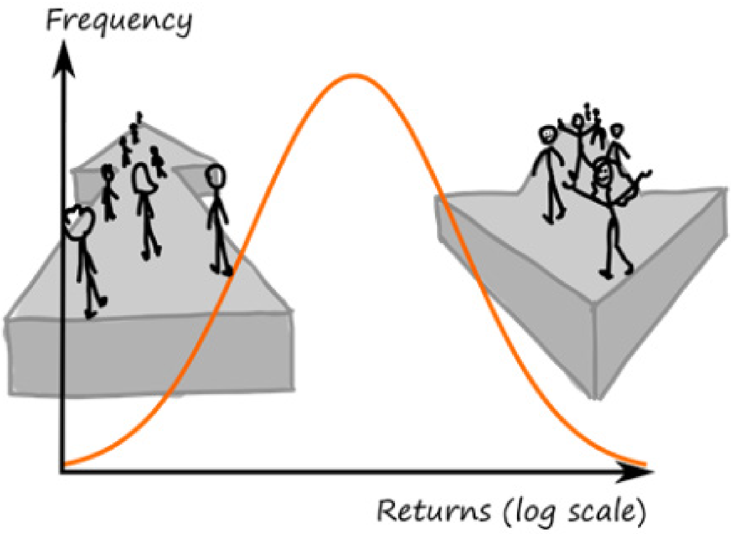


Figure 36: Altruistic market efficiency (2)

And this comes up in just regular markets. Ways to make money, maybe they actually start out distributed across a wide range. This is a log scale now, and it is meant to represent one of those heavy-tailed distributions, but then people who are losing money, say, "Well, this sucks," and they stopped doing that thing. And they see other people who are doing activities which are making lots of money. And they think: "Yeah, I'm going to go do that." And then you get more people going into that area, and then diminishing returns mean that you actually make less money than you used to, by doing stuff in that area. So, afterwards, you end up with a much more narrow distribution of the value that is being produced by people doing these different things, than we started with.

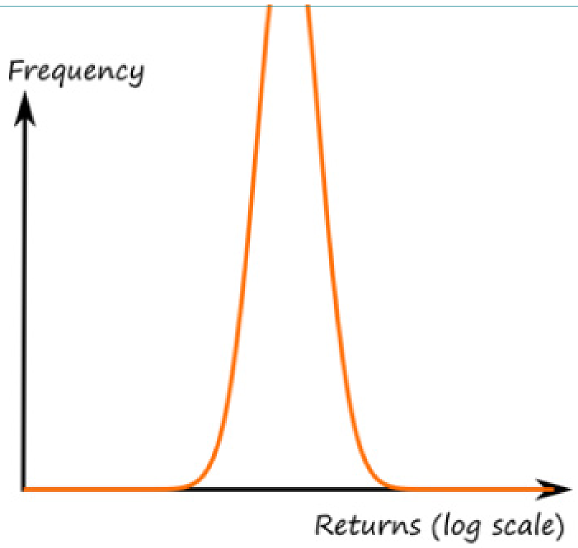


Figure 37: Altruistic market efficiency (3)

We might get a push in that direction among opportunities to create altruistic value. I certainly don't think that we are in a properly efficient market. I'm not sure how efficient it is, how much we are curving that tail. I hope that as this community grows, as we get more people who are actively trying to choose very valuable things, that will mean the distribution does get less heavy-tailed.

One of the mechanisms that lead to efficiency in regular markets is feedback loops, where people just notice they are getting rich or that they are losing money. Another mechanism is people doing analysis, and they do this because of the feedback loops, trying to work out that actually, we should put more resources there because then we'll get richer. I think that doing that analysis is an important part of the project we're collectively embarking on here.

Overall, I don't think that we do have an efficient market for this. I do believe we have heavy-tailed distributions. I'm not sure how extreme, but that's because it responds to actions people are taking.

Factoring cost-effectiveness

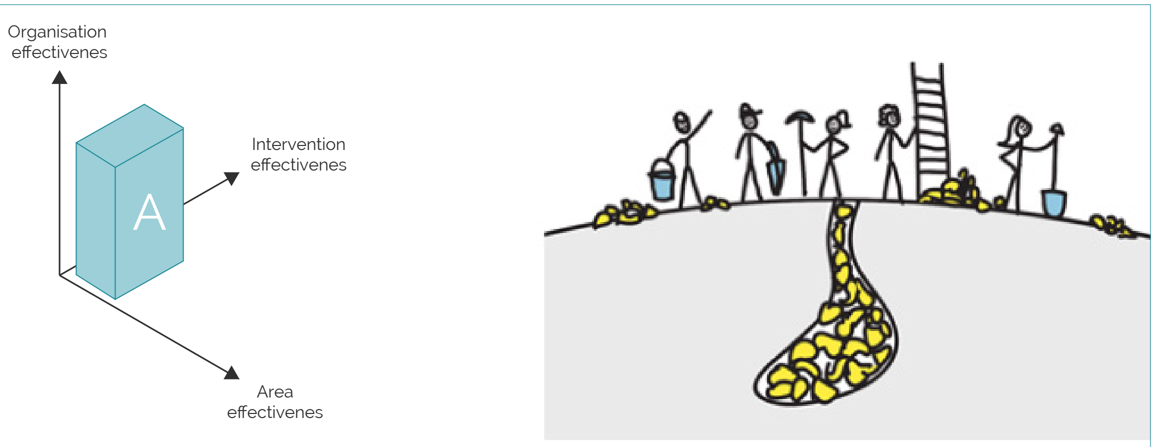


Figure 38: Factoring cost-effectiveness

Factoring cost-effectiveness: I think that this is just an extremely simple point, and there isn't really space for it to be wrong. But there is an empirical question as to how much these different dimensions matter. It might be that you just have way more variation in one of the dimensions than others. Actually, I don't have that much of a view of how much the different dimensions matter. We saw that the intervention effectiveness within global health varied by three or four orders of magnitude. Area effectiveness, I think, may be more than that, but I'm not sure how much more. In terms of organization effectiveness, I'm just not an expert, and I don't want to try and claim to have much of a view on that.

Diminishing returns

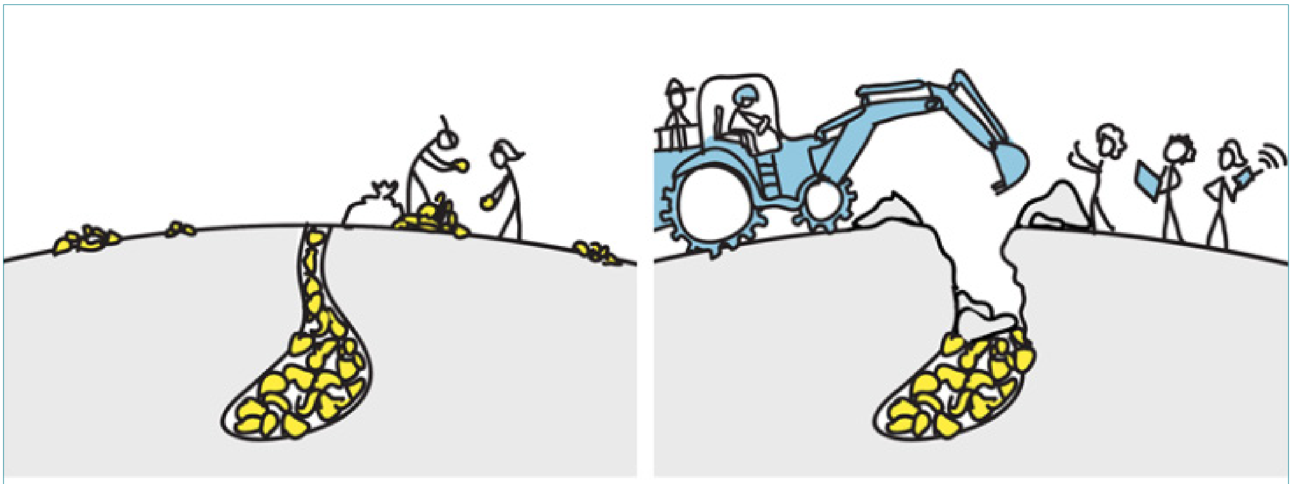


Figure 39: Diminishing returns

Diminishing returns: I just think this is an extremely robust point. Sometimes, in some domains, there are increasing returns to scale, where you get efficiencies of scale, and that helps you. I think that more often applies to the organization scale or organization within a domain. Whereas diminishing returns often apply at the domain scale. But I do know some smart people who think that I am overstating the case for diminishing returns. So although I think, personally, that there's a pretty robust case, I would add a note of caution there.

Scale, tractability, uncrowdedness

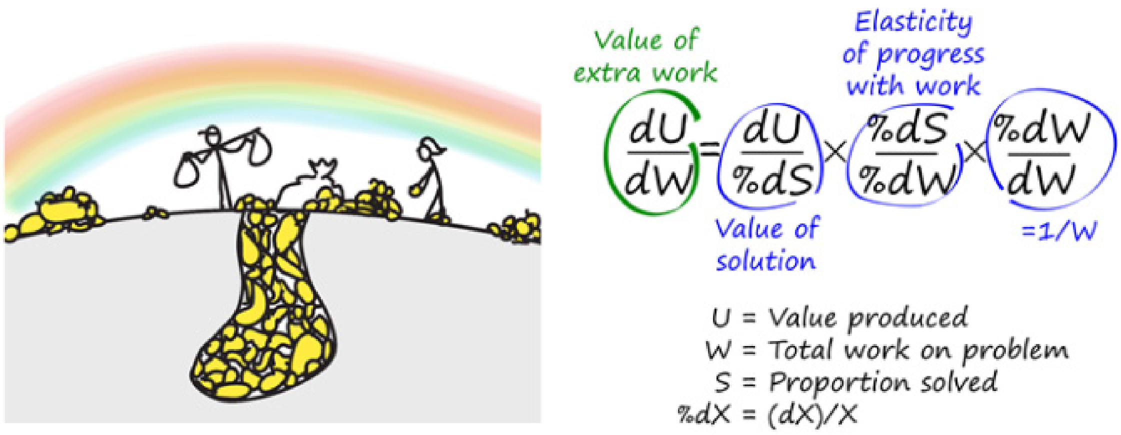

Value of extra work
Elasticity of progress with work
$\frac{dU}{dW} = \frac{dU}{\%dS} \times \frac{\%dS}{\%dW} \times \frac{\%dW}{dW}$
Value of solution
=1/W
U = Value produced
W = Total work on problem
S = Proportion solved
%dX = (dX)/X

Figure 40: Scale, tractability, uncrowdedness

Scale, tractability, neglectedness: I think it is obvious that they all matter. I think it is obvious, it is just trivial, that this factorization is correct as a factorization. What is less clear is whether this breaks it up into things that are easier to measure and whether this is a helpful way of doing it. I think it probably is. We get some evidence from the fact that it loosely matches up with an informal framework that people have been using for a few years, and have seemed to find helpful.

Absolute and marginal priority

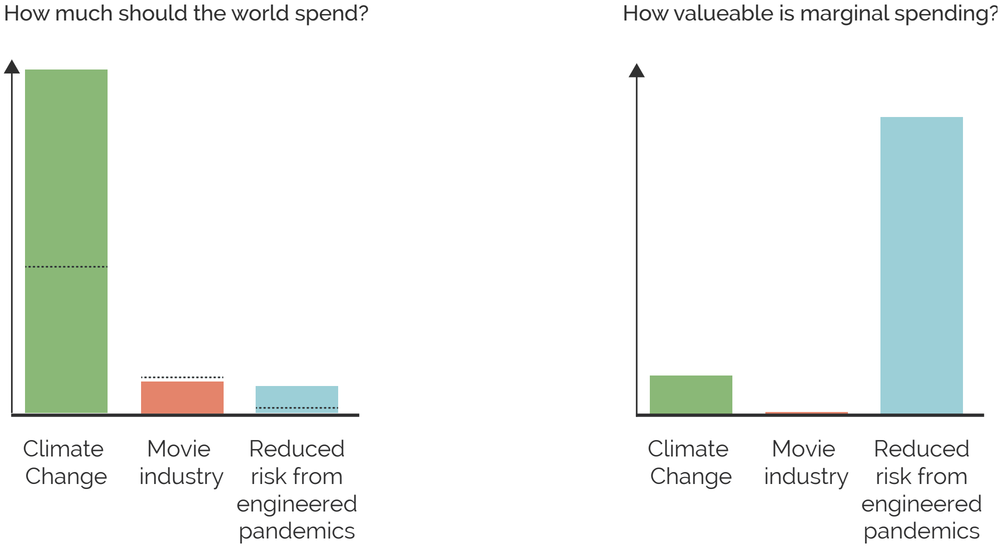


Figure 41: Absolute and marginal priority

Absolute and marginal priority: Again, at some level, is just trivial. I made this point about communication because I think not everybody has these separate notions and we can confuse each other if we blur them.

Differential progress

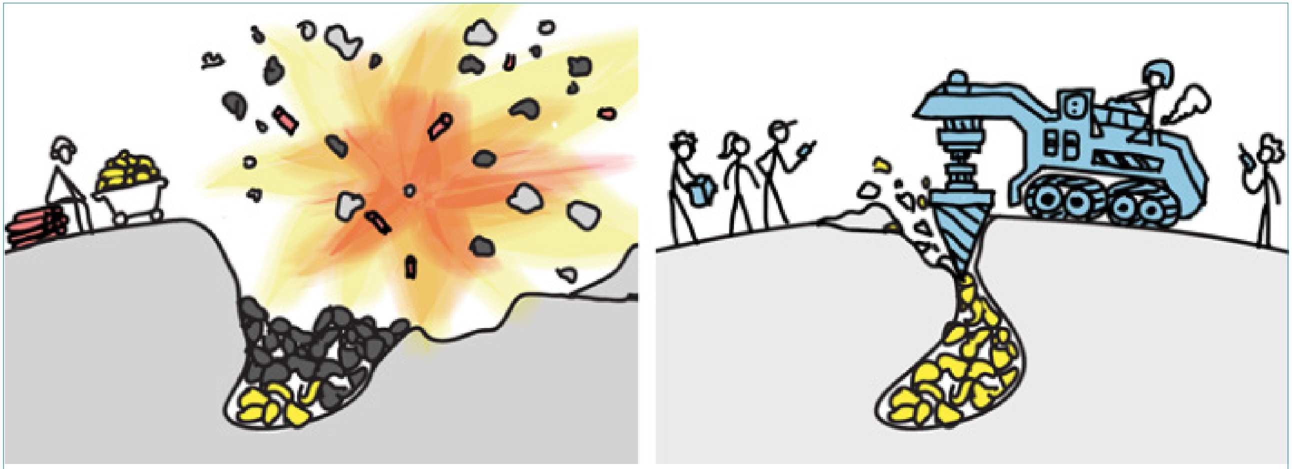


Figure 42: Differential progress

Differential progress: I think that this argument basically checks out. It appears in a few academic papers. It’s also believed by some of the smartest and most reasonable people I know, which gives me some evidence that it might be true, outside of my personal introspection. It hasn’t had that much scrutiny, and it's a bit counterintuitive, so maybe we want to expose it to more scrutiny.

Comparative advantage

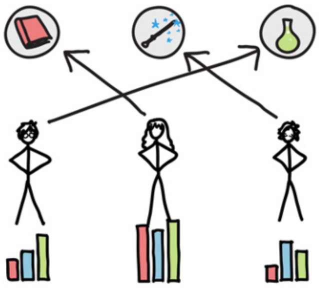


Figure 43: Comparative advantage

Comparative advantage is just a pretty standard idea from economics. Normally markets try to work to push people into working in the way that utilizes their comparative advantage. We don't necessarily have that when we're aiming for more altruistic value.

The application across time is also a bit more speculative. I'm one of the main people who has been trying to reason this way. I haven't had anybody push back on it, but take it with a bit more salt, because it's just less well checked out.

Aggregating knowledge

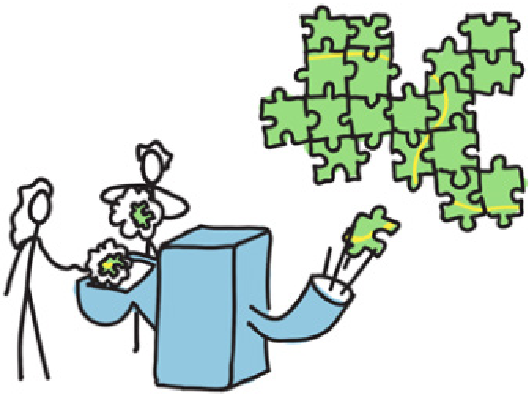


Figure 44: Aggregating knowledge

Aggregating knowledge: I think everyone tends to think that yes, we want intuitions for this. And I think there is also pretty broad consensus that the existing institutions are not perfect. Whether we can build better institutions, I'm less certain about.

Sharing reasons for beliefs

Stating reasons for beliefs: this again is something that I think is common sense. All else equal, this is a good thing. But of course, there are costs to doing it. It slows down our communication. And it may just not sound glamorous and therefore be harder to get people on board with this. I think that at least

Figure 45: Sharing reasons for beliefs

we want to nudge people in this direction, but I don't know exactly how far in this direction. We don't want to be overwhelmingly demanding on this. I, to some extent, believe this because a load of smart, reasonable people I know, think that we want to go in this direction. And I weigh other people's opinions when I don't see a reason that I should have a notably better perspective on it than them.

Conclusion

Finally, why have I been sharing all of this with you? You know, people can go and mine gold without understanding all these theoretical arguments about the distribution of gold in the world. But, because it's invisible, we need to be more careful about aiming at the right things. And so I think it's more important for our community to have this knowledge broadly spread. And I think that we are still in the early days of the community and so it's particularly important to try and get this knowledge in at the foundations and to work out better versions of this. We don't want to have the kind of gold rush phenomenon where people charge off after a thing, and it turns out there wasn't actually that much value there.

Crucial Considerations and Wise Philanthropy

By Nick Bostrom, 2014

This article was originally published as 'Crucial Considerations and Wise Philanthropy' on www.stafforini.com
In this 2014 talk, the Future of Humanity Institute's Nick Bostrom discusses the concept of crucial considerations and how we can use it to maximize our impact on the long-term future. Listen to the talk by clicking on the image below. The below transcript is lightly edited for readability.

Good Done Right
Nick Bostrom: "Crucial considerations and wise philanthropy"
SOUNDCLOUD
Share
Cookie policy

What is a crucial consideration?

I want to talk about this concept of a crucial consideration, which comes up in the work that we're doing a lot. Suppose you're out in the forest and you have a map and a compass, and you're trying to find some destination. You're carrying some weight, maybe you have a lot of water because you need to hydrate yourself to reach your goal and carry weight, and trying to fine-tune the exact direction you're going. You're trying to figure out how much water you can pour out, to lighten your load without having too little to reach your destination.

All of these are normal considerations: you're fine-tuning the way you're going to make more rapid progress towards your goal. But then you look more closely at this compass that you have been using, and you realize that the magnet part has actually come loose. This means that the needle might now be pointing in a completely different direction that bears no relation to North: it might have rotated some unknown number of laps or parts of a lap.

With this discovery, you now completely lose confidence in all the earlier reasoning that was based on trying to get the more accurate reading of where the needle was pointing. This would be an example of a crucial consideration in the context of orienteering. The idea is that there could be similar types of consideration in more important contexts, that throw us off completely. So a crucial consideration is a consideration such that if it were taken into account it would overturn the conclusions we would otherwise reach about how we should direct our efforts, or an idea or argument that might possibly reveal the need not just for some minor course adjustment in our practical endeavors but a major change of direction or priority.

Within a utilitarian context, one can perhaps try to explicate it as follows: a crucial consideration is a consideration that radically changes the expected value of pursuing some high-level subgoal. The idea here is that you have some evaluation standard that is fixed, and you form some overall plan to achieve some high-level subgoal. This is your idea of how to maximize this evaluation standard. A crucial consideration, then, would be a consideration that radically changes the expected value of achieving this subgoal, and we will see some examples of this. Now if you widen the context not limited to some utilitarian context, then you might want to retreat to these earlier more informal formulations, because one of the things that could be questioned is utilitarianism itself. But for most of this talk we will be kind of thinking about that component.

There are some related concepts that are useful to have. So a crucial consideration component will be an argument, idea or datum which, while not on its own amounting to a crucial consideration, seems to have a substantial probability of maybe being able to serve a central role within a crucial consideration. It's the kind of thing of which we would say: "This looks really intriguing, this could be important; I'm not really sure what to make of it at the moment." On its own maybe it doesn't tell us anything, but maybe there's another piece that, when combined, will somehow yield an important result. So those kinds of crucial consideration components could be useful to discover.

Then there's the concept of a deliberation ladder, which would be a sequence of crucial considerations, regarding the same high-level subgoal, where the considerations hold in opposing directions. Let's look at some examples of these kinds of crucial consideration ladders that help to illustrate the general predicament.

Should I vote in the national election?

Let's take this question: "Should I vote in the national election?" At the sort of "level one" of reasoning, you think, "Yes, I should vote to put a better candidate in office." That clearly makes sense.

Then you reflect some more: "But, my vote is extremely unlikely to make a difference. I should not vote, but put my time to better use."

(These examples are meant to illustrate the general idea; it's not so much I want a big discussion as to these particular examples, they're kind of complicated. But I think they will serve to illustrate the general phenomenon.)

So here we have gone from "Yes, we should vote," making a plan to get to the polling booth, etc. And then, with the consideration number two, "No, I should not vote. I should do something completely different."

Then you think, "Well, although it's unlikely that my vote will make a difference, the stakes are very high: millions of lives are affected by the president. So even if the chance that my vote will be decisive is one in several million, the expected benefit is still large enough to be worth a trip to the polling station." So I just went back to the television, turned on the football game, and now it turns out I should vote, so we have a reversed direction.

Then you continue to think, "Well, if the election is not close, then my vote will make no difference. If the election is close, then approximately half of the votes will be for the wrong candidate, implying either that the candidates are exactly or almost exactly of the same merit, so it doesn't really matter who wins, or typical voters' judgment of the candidates' merits is extremely unreliable, and carries almost no signal, so

I should not bother to vote."

Now you sink back into the comfy sofa and bring out the popcorn or whatever, and then you think, "Oh, well, of course I'm a much better judge of the candidates' merits than the typical voter, so I should vote."

Then you think, "Well, but psychological studies show that people who tend to be overconfident almost everybody believes themselves to be above average, but they are as likely to be wrong as right about that. If I am as likely to vote for the wrong candidate as is the typical voter, then my vote would have negligible information to the selection process, and I should not vote."

Then we go on...

"Okay, I've gone through all of this reasoning that really means that I'm special, so I should vote."

But then, "Well, if I'm so special, then the opportunity cost..." (This is why I warned you all against becoming philosophers.)

So, I should do something more important. But if I don't vote my acquaintances will see that I have failed to support the candidates that we all think are best, they would think me weird and strange, and disloyal. Then that would maybe diminish my influence, which I could otherwise have used for good ends, so I should vote after all.

But it's important to stand up for one's convictions, to stimulate fruitful discussion. They might think me like really sophisticated if I explained all this complicated reasoning for voting, and that might increase my influence, which I can then invest in some good cause. Etc, etc, etc.

There is no reason to think that the ladder would stop there; it's just that we ran out of steam at this point. If you end at some point, you might then wonder, maybe there are further steps on the ladder, and how much reason do you really think you have for the conclusion you're temporarily at, at that stage?

Should we favor more funding for x-risk tech research?

I want to look at one other example of a deliberation ladder more in the context of technology policy and x-risk. This is a kind of argument that can be run with regard to certain types of technologies, whether we should try to promote them or get more funding from.

The technology here is nanotechnology—this is in fact the example where this line of reasoning originally came up. Some parts of this hark back to Eric Drexler's book Engines of Creation, where he actually advocated this line of thinking (ch. 12). So we should fund nanotechnology — this is the "level one" reasoning — because there are many potential future applications: medicine, manufacturing, clean energy, etc. It would be really great if we had all those benefits.

But it also looks like nanotechnology could have important military applications, and it could be used by terrorists etc., to create new weapons of mass destruction that could pose a major existential threat. If it's so dangerous, no, maybe we shouldn't really fund it.

But if this kind of technology is possible, it will almost certainly be developed sooner or later, whether or not we decide to pursue it. (‘We’ being maybe the people in this room or the people in Britain or Western democracies.) If responsible people refrain from developing it, then it will be developed by irresponsible people, which would make the risks even greater, so we should fund it.

(You can see that the same template could be relevant for evaluating other technologies with upsides and downsides, besides nanotechnology.)

But we are already ahead in its development, so extra funding would only get us there sooner, leaving us less time to prepare for the dangers. So we should not add funding: the responsible people can get there first even without adding funding to this endeavor.

But then you look around and see virtually no serious effort to prepare for the dangers of nanotechnology — and this is basically Drexler’s point back in Engines —, because serious preparation will begin only after a massive project is already underway to develop nanotechnology. Only then will people take the prospect seriously.

The earlier a serious Manhattan-like project to develop nanotechnology is initiated, the longer it will take to complete, because the earlier you start, the lower the foundation from which you begin. The actual project will then run for longer, and that will then mean more time for preparation: serious preparation only starts when the project starts, and the sooner the project starts, the longer it will take, so the longer the preparation time will be. And that suggests that we should push as hard as we can to get this product launched immediately, to maximize time for preparation.

But there are more considerations that should be taken into account. The level of risk will be affected by factors other than the amount of serious preparation that has been made, specifically to counter the threat from nanotechnology. For instance, machine intelligence or ubiquitous surveillance might be developed before nanotechnology, eliminating or mitigating the risks of the latter. Although these other technologies may pose great risks of their own, those risks would have to be faced anyway. And there’s a lot more that can be said.

Nanotechnology would not really reduce these other risks, like the risks from AI, for example. The preferred sequence is that we get superintelligence or ubiquitous surveillance before nanotechnology, and so we should oppose extra funding for nanotechnology even though superintelligence and ubiquitous surveillance might be very dangerous on their own, including posing existential risk, given certain background assumptions about the technological completion conjecture — that in the fullness of time, unless civilization collapses, all possible general useful technologies will be developed —, these dangers will have to be confronted, and all our choice really concerns is the sequence in which we confront them. And it’s better to confront superintelligence before nanotechnology because superintelligence can obviate the nanotechnology risk, but not vice versa.

However, if people oppose extra funding for nanotechnology, then people working in nanotechnology will dislike those people who are opposing it. (This is also a point from Drexler’s book.) But other scientists might regard these people who oppose funding for nanotechnology as being anti-science and this will reduce our ability to work with these scientists, hampering our efforts on more specific issues—efforts that stand a better chance of making a material difference to any attempt on our part to influence the level of national funding for nanotechnology. So we should not oppose nanotechnology. That is, rather than opposing nanotechnology — we may try to slow it down a little bit. But we are a small group and we can't make a big difference —, we should work with the nanotechnology scientists, be their friend, and then maybe try to influence on the margin, so that they develop nanotechnology in a slightly different way or add some safeguards, and stuff like that.

Again, there is no clear reason to think that we have reached the limit of the level of deliberation that we could apply to this. It's disconcerting because it looks like the practical upshot keeps switching back and forth as we look more deeply into the search tree, and we might wonder why this is so. I think that these deliberation ladders are particularly likely to turn up when one is trying to be a thoroughgoing utilitarian and one really takes the big-picture question seriously.

Crucial considerations and utilitarianism

Let's consider some possible reasons for why that might be. If we compare, for example, the domain of application of utilitarianism to another domain of application, say if you have an ordinary human preference function—you want a flourishing life, like a healthy family, a successful career and some relaxation, like a typical human values — if you're trying to satisfy those, it looks less likely that you will encounter a large number of these crucial considerations. Why might that be?

One possible explanation is that we have more knowledge and experience of human life at the personal level. Billions of people have tried to maximize an ordinary human utility function and have received a lot of feedback and a lot of things have been tried out. So we already know some of the basics like, if you want to go on for decades, it's a good idea to eat, things like that.

They're not something we need to discover. And maybe our preferences in the first place have been shaped to more or less fit the kind of opportunities we can cognitively exploit in the environment by evolution. So we might not have some weird preference that there was no way that we could systematically satisfy. Whereas with utilitarianism, the utilitarian preference extends far and wide beyond our familiar environment, including into the cosmic commons and billions of years into the future and super advanced civilizations: what they do matters from the utilitarian perspective, and matters a lot. Most of what the utilitarian preference cares about is stuff that we have no familiarity with.

Another possible source of crucial considerations with regard to utilitarianism is difficulties in understanding the goal itself. For example, if one tries to think about how to apply utilitarianism to a world that has a finite probability of being infinite, one will run into difficulties in terms of how to measure different infinite magnitudes and still seeing how we could possibly make any difference to it. I have a big paper about that and we don't need to go into that. There are some other issues that consist in actually trying to articulate utilitarianism to deal with all these possible cases.

The third possible reason here is that one might think that we are kind of close, not super close, but close to some pivot point in history. That means that we might have special opportunities to influence the long-term future now. And we're still far enough away from this: it's not obvious what we should do to have the maximally beneficial impact on the future. But still close enough that we can maybe begin to perceive some contours of the apparatus that will shape the future. For example, you may think that superintelligence might be this pivot point, or one of them (there may be x-risk pivot points as well), that we will confront in this century, then it might just be that we are barely just beginning to get the ability to think about those things, which introduces a whole set of new considerations that might be very important.

This could affect the personal domain as well. It's just like with an ordinary person's typical utility function: they probably don't place a million times more value on living for a billion years than living for a hundred years, or a thousand times more value on raising a thousand children than on raising one child. So even though the future still exists, it just doesn't weigh as heavily in a normal human utility function as it does for utilitarians.

Fourthly, one might also argue that we have recently discovered some key exploration tools that enable us to make these very important discoveries about how to be a good utilitarian. And we haven't yet run the course with these tools, so we keep turning up like fundamental new important discoveries using these exploration tools. That's why there seem to be so many crucial considerations being discovered. We might talk a little bit about some of those later in the presentation.

Evaluation functions

Now let me come at this from a slightly different angle. In chess, the way you would ideally play is you would start by thinking the possible moves that you could make, then the possible responses that your component could make, and your responses to those responses. Ideally, you would think that through all the way to the end state, and then just try to select a first move that would be best from the point of view of winning when it could calculate through the entire game tree. But that's computational infeasible because the tree branch is too much: you have an exponential number of moves to consider.

So what you instead have to do is to calculate explicitly some number of plies ahead. Maybe a dozen plies ahead or something like that. At that point, your analysis has to stop, and what you do is to have some evaluation function which is relatively simple to compute, which tries to look at the board state that could result from this sequence of six moves and countermoves, and in some rough and ready way try to estimate how good that state is. A typical chess evaluation function might look something like this.
You have some term that evaluates how much material we have, like having your queen and a lot of pieces is beneficial. The opponent having few of those is also beneficial. We have some metric like a pawn is

$\text{Eval}_{\text{chess}} = (c_1 \times \text{material}) + (c_2 \times \text{mobility}) + (c_3 \times \text{king safety}) + (c_4 \times \text{ center control}) + \ldots$

worth one and queen is worth, I don't know, 11 or something like that.

So you weigh that up — that's one component in the evaluation function. Then maybe consider how mobile your pieces are. If they're all crammed in the corner, that's usually an unpromising situation, so you have some term for that. King safety - center control adds a bit of value: if you control the middle of the board, we know from experience that tends to a good position.

So what you do is calculate explicitly some number of steps ahead and then you have this relatively unchanging evaluation function that is used to figure out which of these initial games that you could play would be resulting in the most beneficial situation for you. These evaluation functions are mainly derived from some human chess masters who have a lot of experience playing the game. The parameters, like the weight you assign to these different features, might also be learned by machine intelligence.

We do something analogous to that in other domains. Like a typical traditional public policy, social welfare economists might think that you need to maximize some social welfare function which might take a form like this.

GDP? Yes, we want more GDP, but we also have to take into account the amount of unemployment, maybe the amount of equality or inequality, some factor for the health of the environment. It might not be that whatever we write there s exactly the thing that is equivalent to moral goodness fundamentally considered. But we know that these things tend to be good, or we think so.

This is a useful approximation of true value that might be more tractable in a practical decision-making context. One thing I can ask, then, is if there is something similar to that for moral goodness. You want to do the morally best thing you can do, but to calculate all of these out from scratch just looks difficult or impossible to do in any one situation. You need more stable principles that you can use to evaluate different things you could do. Here we might look at the more restricted version of utilitarianism. We can wonder what we might put in there.

Evaluation function

$$\text{Eval}_{\text{chess}} = (c_1 \times \text{material}) + (c_2 \times \text{mobility}) + (c_3 \times \text{king safety}) + (c_4 \times \text{center control}) + \ldots$$

$$\text{Eval}_{\text{public\_policy}} = (c_1 \times \text{GDP}) + (c_2 \times \text{employment}) + (c_3 \times \text{equality}) + (c_4 \times \text{environment}) + \ldots$$

Here we can hark back to some of the things Beckstead talked about. If we plot capacity, which could be level of economic development and technological sophistication, stuff like that, on one axis and time on the other, my view is that the human condition is a kind of metastable region on this capability axis. You might fluctuate inside for a while, but the longer the time scale you're considering, the greater the chance that you will exit that region in either the downwards direction and go extinct—if you have too few resources below the minimum viable population size, you go extinct (that's one attractor state: once you're extinct, you tend to stay extinct) — or in the upwards direction: we get through to technological

Evaluation function

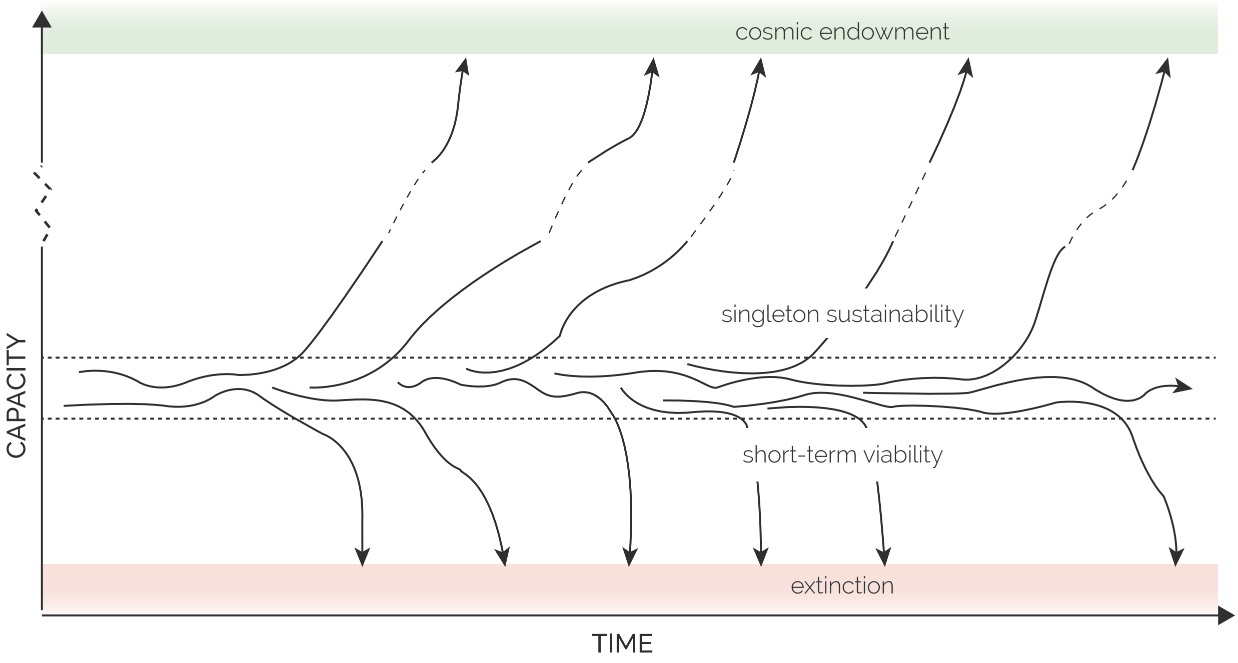
$$\mathrm{Eval}_{chess} = (c_1 \times \text{material}) + (c_2 \times \text{mobility}) + (c_3 \times \text{king safety}) + (c_4 \times \text{center control}) + \ldots$$

$$\mathrm{Eval}_{public\_policy} = (c_1 \times \text{GDP}) + (c_2 \times \text{employment}) + (c_3 \times \text{equality}) + (c_4 \times \text{environment}) + \ldots$$

$$\mathrm{Eval}_{utilitarian} = (c_1 \times \text{GDP}) + (c_2 \times \text{employment}) + (c_3 \times \text{equality}) + (c_4 \times \text{environment}) + \ldots$$

maturity, start colonization process and the future of earth-originating intelligent life might just then be this bubble that expands at some significant fraction of the speed of light and eventually accesses all the cosmological resources that are in principle accessible from our starting point. It's a finite quantity because of the positive cosmological constant: looks like we can only access a finite amount of stuff. But once you've started that, once you're an intergalactic empire, it looks like it could just keep going with high probability to this natural vision.

We can define the concept of an existential risk as one that fails to realize the potential for value that you could gain by accessing the cosmological commons, either by going extinct or by maybe accessing all the cosmological commons but then failing to use them for beneficial purposes or something like that.



That suggests this Maxipok principle that Beckstead also mentioned: Maximize the probability of an OK outcome. That's clearly, at best, a rule of thumb: it's not meant to be a valid moral principle that's true in all possible situations. It's not that. In fact, if you want to go away from the original principle you started with to something practically tractable, you have to make it contingent on various empirical assumptions.

That's the trade-off there: you want to make as weak assumptions as you can and still move it as far as possible towards being tractable as you can.

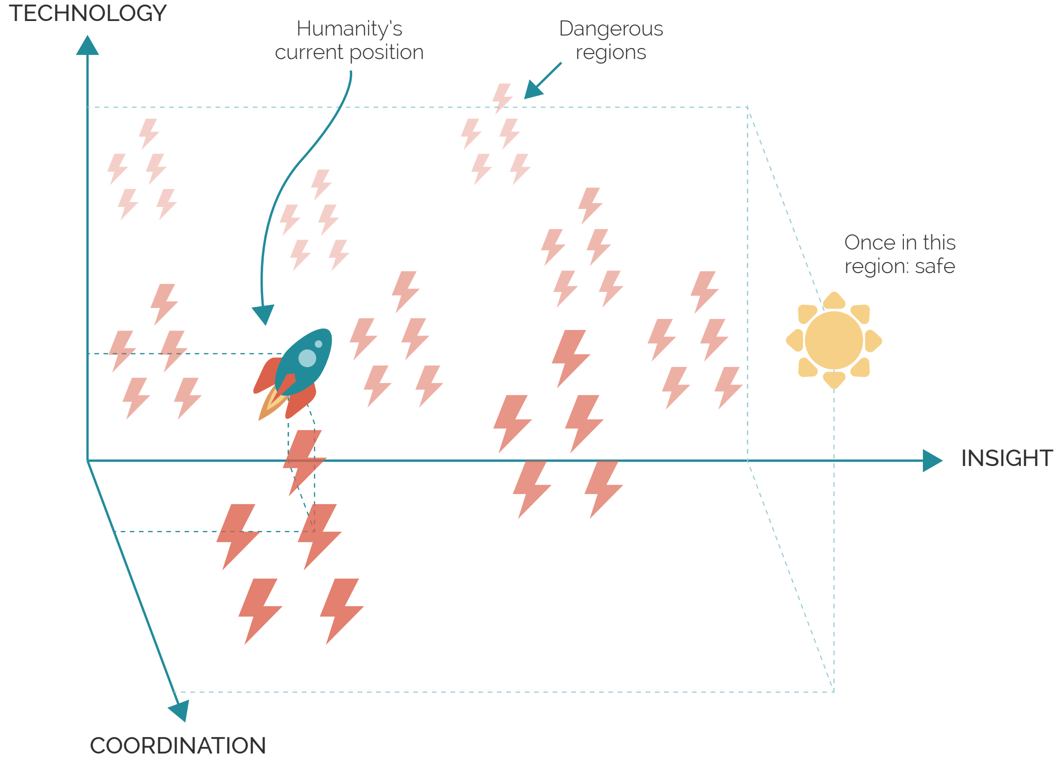
I think this is something that makes a reasonable compromise there. In other words, take the action that minimizes the integral of existential risk that humanity will confront. It will not always give you the right answer, but it's a starting point. There are different things to the ones that Beckstead mentioned, there could be other scenarios where this would give the wrong answer: if you thought that there was a big risk of hyper existential catastrophe like some health scenario, then you might want to increase level of existential risks slightly in order to decrease the risk that there would not just be an existential catastrophe but hyper existential catastrophe. Other things that could come into it are trajectory changes that are less than drastic and just shift slightly.

For present purposes, we could consider the suggestion of using the Maxipok rule as our attempt to define the value function for utilitarian agents. Then the question becomes, If you want to minimize existential risk, what should you do? That is still a very high-level objective. We still need to do more work to break that down into more tangible components.

I'm not sure how well this fits in with the rest of the presentation.

I have this nice slide from another presentation. It's a different way of saying some of what I just said: instead of thinking about sustainability as is commonly known, as this static concept that has a stable state that we should try to approximate, where we use up no more resources than are regenerated by the natural environment, we need, I think, to think about sustainability in dynamical terms, where instead of

Dynamic sustainability



reaching a state, we try to enter and stay on a trajectory that is indefinitely sustainable in the sense that we can contain it to travel on that trajectory indefinitely and it leads in a good direction.

An analogy here would be if you have a rocket. One stable state for a rocket is on the launch pad: it can stand there for a long time. Another stable state is if it's up in space, it can continue to travel for an even longer time, perhaps, if it doesn't rust and stuff. But in mid-air, you have this unstable system. I think that's where humanity is now: we're in mid-air. The static sustainability concept suggests that we should reduce our fuel consumption to the minimum that just enables us to hover there. Thus, maybe prolong the duration in which we could stay in our current situation, but what we perhaps instead should do is maximize the fuel consumption so that we have enough thrust to reach escape velocity. (And that's not a literal argument for burning as much fossil fuel as possible. It's just a metaphor.)

The point here is that to have the best possible condition; we need super advanced technology: to be able to access the cosmic commons, to be able to cure all the diseases that plague us, etc. I think to have the best possible world, you'll also need a huge amount of insight and wisdom, and a large amount of coordination so as to avoid using high technology to wage war against one another, and so forth. Ultimately, we would want a state where we have huge quantities of each of these three variables, but that leaves open the question of what we want more from our consideration. It might be, for example, that we would want more coordination and insight before we have more technology of a certain type. So that before we have various powerful technologies, we would first want to make sure that we have enough peace and understanding to not use them for warfare, and that we have enough insight and wisdom not to accidentally blow ourselves up with them.

A superintelligence, clearly, seems to be something you want in utopia — it's a very high level of technology — but we might want a certain amount of insight before we develop superintelligence, so we can develop it in the correct way.

One can begin to think about, as in analogy with the computer test situation, if there are different features that one could possibly think of as components of this evaluation function for the utilitarian, the Maxipok.

This principle of differential technological development suggests that we should retard the development of dangerous and harmful technologies—once the raise existential risk, that is— and accelerate technologies that reduce existential risks.

Here is our first sketch, this is not a final answer, but one may think that we want a lot of wisdom, we want a lot of international peace and cooperation, and with regards to technologies, it gets a little bit more complicated: we want faster progress in some technology areas, perhaps, and slower in others. I think those are three broad kinds of things one might want to put into the one's evaluation function.

This suggests that one thing to be thinking about in addition to interventions or causes, is the signature of different kinds of things. An intervention should be sort of high leverage, and cause area should promise high leverage interventions. It's not enough that something you could do would do good, you also want to think hard about how much good it could do relative to other things you could do. There is no point in thinking about causes without thinking about how do you see all the low hanging fruit that you could access. So a lot of the thinking is about that.

But when we're moving at this more elevated plane, this high altitude where there are these crucial considerations, then it also seems to become valuable to think about determining the sign of different basic parameters, maybe even where we are not sure how we could affect them. (The sign being, basically, Do we want more or less of it?) We might initially bracket questions as to leverage here, because to first orient ourselves in the landscape we might want sort of postpone that question a little bit in this context. But a good signpost — that is a good parameter of which we would like to determine the signature — would have to be visible from afar. That is, if we define some quantity in terms that still make it very difficult for any particular intervention to say whether it contributes positively or negatively to this quantity that we just defined, then it's not so useful as a signpost. So, "maximize expected value", say, is the quantity they could define. It just doesn't help us very much, because whenever you try to do something specific you're still virtually as far away as you had been. On the other hand, if you set some more concrete objective, like maximize the number of people in this room, or something like that, we can now easily tell like how many people there are, and we have ideas about how we could maximize it. So any particular action we think of we might easily see how it fares on this objective of maximizing the people in this room. However, we might feel it's very difficult to get strong reasons for knowing whether more people in this room is better, or whether there is some inverse relationship. A good signpost would strike a reasonable compromise between being visible from afar and also being such that we can have strong reason to be sure of its sign.

Some tentative signposts

Here are some very tentative signposts: they're tentative in my own view, and I guess there might also be a lot of disagreement among different people. So these are more like areas for investigation. But it might be useful just to show how one might begin to think about it.

Do we want faster progress in computer hardware or slower progress? My best guess there is that we want slower progress. And that has to do with the risks from the machine intelligence transition. Faster computers would make it easier to make AI, which (a) would make them happen sooner probably, which seems perhaps in itself because it leaves less time for the relevant kind of preparation, of which there is a great need; and (b) might reduce the skill level that would be required to produce AI: with a ridiculously large amount of computing power you might be able to produce AI without really knowing much about what you're doing; when you are hardware-constrained you might need more insight and understanding, and it's better that AI be created by people who have more insight and understanding.

This is not by any means a knockdown argument, because there are other existential risks. If you thought that we are about to go extinct anytime soon, because somebody will develop nanotechnology, then you might want to sort of try the AI wildcard as soon as possible. But all-things-considered this is my current best guess. These are the kinds of reasoning that one can engage in.

Whole brain emulation? We did a long, big analysis of that. More specifically, not whether we want to have whole brain emulation, but whether we want to have more or less funding for whole brain emulation, more or fewer resources for developing that. This is one possible path towards machine superintelligence, and for complicated reasons, my guess is "No", but that's even more uncertain, and we have a lot of

Some (very) tentative signposts

Computer hardware?	No
Whole brain emulation?	No (?)
Biological cognitive enhancement?	Yes
Artificial intelligence?	No
Lead of AI frontrunner?	Yes
Solutions to the control problem?	Yes
Effective altruism movement?	Yes
International peace and cooperation?	Yes
Synthetic biology?	No(?)
Nanotechnology?	No
Economic growth?	?
Small and medium-scale catastrophe prevention?	?

different views in our research group on that. (In the discussion, if anybody is interested in one particular one, we can zoom in on that.)

Biological cognitive enhancement of humans? My best guess there is that we want faster progress in that area.

So with these three — I talk more about them in the book — and AI as well.

AI I think we want AI probably to happen a little bit slower than it's likely to do by default.

Another question is: If there is one company or project or team that will develop the first successful AI, how much ahead does one want that team to be to the second team that is trying to do it? My best guess is that we want it to have a lot of lead, many years ideally, to enable them to slow down at the end to implement more safety measures, rather than being in the tight tech race.

Solutions to the control problem for AI? I think we want faster progress in that, and that's one of our focus areas, and some of our friends from the Machine Intelligence Research Institute are here, also working hard on that.

The effective altruism movement? I think that looks very good in many ways, robustly good, to have faster, better growth in that.

International peace and cooperation? Looks good.

Synthetic biology? I think it looks bad. We haven't thought as carefully about that, so that could change, but it looks like there could be x-risks from that, although it may also be beneficial. Insofar as it might enable improvements in cognitive enhancement, there'll be a kind of difficult trade-off.

Nanotechnology? I think it looks bad: we want slower progress towards that.

Economic growth? Very difficult to tell the sign of that, in my view. And within a community of people have thought hard about that are, again, different guesses as to the sign of that.

Small and medium scale catastrophe prevention? Also looks good. So global catastrophic risks falling short of existential risk. Again, very difficult to know the sign of that. Here we are bracketing leverage at all, even just knowing whether we would want more or less, if we could get it for free, it's non-obvious. On the one hand, small-scale catastrophes might create an immune response that makes us better, puts in place better safeguards, and stuff like that, that could protect us from the big stuff. If we're thinking about medium-scale catastrophes that could cause civilizational collapse, large by ordinary standards but only medium-scale in comparison to existential catastrophes, which are large in this context, again, it is not totally obvious what the sign of that is: there's a lot more work to be done to try to figure that out. If recovery looks very likely, you might then have guesses as to whether the recovered civilization would be more likely to avoid existential catastrophe having gone through this experience or not.

So these are the parameters that one can begin to think about. One doesn't realize just how difficult it is, even some parameters that from an ordinary common-sense point of view seem kind of obvious, actually turn out to be quite non-obvious once you start to think through the way that they're all supposed to fit together.

Suppose you're an administrator here in Oxford, you're working in the Computer Science department, and you're the secretary there. Suppose you find some way to make the department run slightly more efficiently: you create this mailing list so that everybody can, when they have an announcement to make, just email it to the mailing list rather than having to put in each person individually in the address field. And that's a useful thing, that's a great thing: it didn't cost anything, other than one-off cost, and now everybody can go about their business more easily. From this perspective, it's very non-obvious whether that is, in fact, a good thing. It might be contributing to AI—that might be the main effect of this, other than the very small general effect on economic growth. And it might probably be that you have made the world worse in expectation by making this little efficiency improvement. So this project of trying to think through this it's in a sense a little bit like the Nietzschean Umwertung aller Werte — the revaluation of all values—project that he never had a chance to complete, because he went mad before.

Possible areas with additional crucial considerations

So, these are some kinds of areas—I'm not going to go into all of these, I'm just giving examples of the kinds of areas where today it looks like there might still be crucial considerations. This is not an exhaustive list by any means, and we can talk more about some of those. They kind of go from more general and abstract and powerful, to more specific and understandable by ordinary reasoning.
To just pick an example: insects. If you are a classical utilitarian, this consideration arises within the more mundane—we're setting aside the cosmological commons and just thinking about here on Earth. If insects are sentient then maybe the amount of sentience in insects is very large because there are very,

List of some areas with candidate remaining CCs or CCCs

- Counterfactual trade
- Simulation Stuff
- Infinite paralysis
- Pascalian muggings
- Different kinds of aggregative ethics (total, average, negative)
- Information hazards

- Aliens
- Baby universes
- Other kinds of moral uncertainty
- Other game theory stuff

`

- Pessimistic metainduction; epistemic humility; anthropics
- Insects, subroutines

very many of them. So that maybe the effect of our policies on insect well-being mind trump the effect of our policies on human well-being or animals in factories and stuff like that. I'm not saying it does, but it's a question that is non-obvious and that could have a big impact.

Or take another example:

Subroutines. With certain kinds of machine intelligence there are processes, like reinforcement learning algorithms and other subprocesses within the AI, that could turn out to have moral status in some way. Maybe there will be hugely large numbers of runs of these subprocesses, so that if it turns out that some of these kinds of things count for something, then maybe the numbers again would come to dominate.

Some partial remedies

Each of these is a whole workshop on its own, so it's not something we can go into. But what can one do if one suspects that there might be these crucial considerations, some of them not yet discovered? I don't have a crisp answer to that. Here are some prima facie plausible things one might try to do a little bit of:

- *Don't act precipitously*, particularly in ways that are irrevocable.
- *Invest in more analysis* to find and assemble missing crucial considerations. That's why I'm doing the kind of work that I'm doing, and the rest of us are also involved in that enterprise.
- *Take into account that expected value changes are probably smaller than they appear*. If you are a utilitarian, let's say you think of this new argument that has this radical implication for what you should be doing, the first instinct might be to radically change your expected utility of different practical policies in light of this new insight. But maybe when you reflect on the fact that there are new crucial considerations being discovered every once in awhile, maybe you should still change

your expected value, but not as much as it seems you should the first time. You should reflect on this at the meta level.

- *Take into account fundamental moral uncertainty*. If we widen our purview to not just consider utilitarianism, as we should consider things from a more general unrestricted normative perspective, then something like the Parliamentary Model for taking normative uncertainty into account looks fairly robust. This is the idea that if you are unsure as to which moral theory is true, then you should assign probabilities to different moral theories and imagine that there were a parliament where each moral theory got to send delegates to that parliament in proportion to their probability. Then in this imaginary parliament, these delegates from the different moral theories discuss and compromise and work out what to do. And then you should do that what that moral parliament of yours would have decided, as a sort of metaphor. The idea is that, other things equal, the more probability a moral theory has, the greater its say in determining your actions, but there might also be these trades between different moral theories which I think Toby Ord talked about in his presentation. This is one metaphor for how to conceive of those traits. It might not be exactly the right way to think about fundamental normative uncertainty, but it seems to be close in many situations, and it seems to be relatively robust in the sense of being unlikely to have a totally crazy implication.
- *Focus more on near-term and convenient objectives.* To the extent that one is despairing about having any coherent view about how to go about maximizing aggregative welfare in this cosmological context, the greater it seems the effective voice of other types of things that one might be placing weight. So if you're partly an egoist and partly an altruist, then if you say that the altruistic component is on this kind of deliberation ladder then maybe you should go more with the egoistic part, until and unless you can find stability in your altruistic deliberations.
- *Focus on developing our capacity as a civilization to wisely deliberate on these types of things*. To build up our capacity, rather than pursuing very specific goals, and by capacity in this context it looks like perhaps we should focus less on powers and more on the propensity to use powers as well. This is still quite vague, but something in that general direction seems to be robustly desirable. Certainly, you could have a crucial consideration that's turned up to show that that was the wrong thing to do, but it still looks like a reasonable guess.

That's it. Thanks.

The Moral Value of Information

Amanda Askell, 2017

In this 2017 talk, the NYU philosopher Amanda Askell argues that we often underestimate the value of new information or knowledge when thinking of how to do good. This means that interventions which give us information, such as research, are often more valuable than it might naively seem. The below transcript is edited for readability.

I'm going to start the talk with some spoilers. Basically, I want nothing here to be a surprise. The first thing that I'm going to claim, and I hope you find it plausible, is that we generally prefer interventions with more evidential support, all else being equal. I'll go into detail about what that means. The second is, I'm going to argue that having less evidence in favor of a given intervention means that your credences about the effectiveness of that intervention are what I call "low resilience".

This is something that has been explored in decision theory to some extent. That's true even if your credences about the effectiveness of that intervention are the same value. So, if I thought there was a 50% chance that I would get $100, there's actually a difference between a low resilience 50% and a high resilience 50%.

I'm going to argue that, if your credences are low resilience, then the value of information in this domain is generally higher than it would be in a domain where your credences are high resilience. And, I'm going to argue that this means that actually in many cases, we should prefer interventions with less evidential support, all else being equal. Hopefully, you'll find that counterintuitive and interesting.

The first thing to say is that we generally think that expected value calculations are a pretty decent way of estimating the effectiveness of a given intervention. An example here is one where we imagine that there is a Disease A, very novelly and interestingly named, and another disease equally interestingly named Disease B [Figure 1].

Basically, the idea is that these two diseases are virtually impossible to differentiate. They all have the same symptoms, they cause the same reduction in life expectancy, etc. The key difference is that they respond very differently to different treatments, so any doctor who finds themselves with a patient with one of these conditions is in a difficult situation.

Expected value and evidence

	Cost ($)	Disease A (0.5)	Disease B (0.5)	Expected years/$
Drug A	100	10 years	0 years	0.05
Drug B	100	0 years	10 years	0.05
Drug C	100	6 years	6 years	0.06

Figure 1

They can prescribe Drug A. Drug A costs $100. If the patient has Disease A, then Drug A will basically extend their life by another 10 years. If on the other hand they have Disease B, it won't extend their life at all. They will die of Disease B, because Disease B is completely non-responsive to Drug A. So the expected years of life that we get from Drug A is 0.05 per dollar. Drug B works in a very similar way, except it is used to treat Disease B. If you have Disease A, it will be completely non-responsive. So, it's got the same expected value.

Then, we have Drug C. Drug C costs $100, but regardless of whether you have Disease A or Disease B, it will in fact be responsive to Drug C. So, this is a new and interesting drug. This means that the expected value for Drug C is greater than the expected value for either Drug A or Drug B. So we think, "Okay, great. Kind of obvious that you should prescribe Drug C."

Suppose that Drug A and Drug B have been heavily tested in numerous trials, and they've been shown in meta-analyses to be highly effective, and that the estimates I just gave you were extremely robust. Drug C on the hand, is completely new. It's only had a single trial, in which it increased patients' lives by many years. We assume that in the trial, this was a trial of patients with both diseases.

So, you have a conservative prior about the effectiveness of a drug. You think, "In likelihood, most random drugs that we were to select would either be net neutral or net negative," so that's your conservative prior. If you see one trial in which a drug massively extends someone's life, then your prior might bring you down to something like six years, regardless of whether they have Disease A or Disease B. Now we have the same expectation, but suddenly it seems a bit more questionable whether we should prescribe Drug C.

This idea that we should favor interventions with more evidence, and that expected utility theory can't capture this, is summed up in this blog post from GiveWell, I think from a couple of years ago.

"There seems to be nothing in explicit expected value that penalizes relative ignorance, or relatively pearly grounded estimates. If I can literally save a child I see drowning by ruining a $1,000 suit, but in the same moment that I make a wild guess that this $1,000 could save two lives if I put it toward medical research, then explicit expected value seems to indicate that I should opt for the latter."

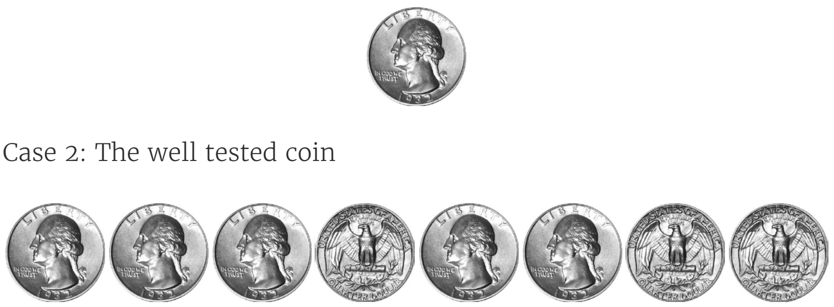
The idea is that there's something wrong with expected value calculations because they kind of tell us to take wild guesses, as long as the expected value is higher. I want to argue that there are kind of two claims that we might want to vindicate in these sorts of cases. The first claim is one that I think I and hopefully you find quite plausible, and it's the claim that evidence matters. So, how much evidence we have about an intervention can make a different to what we should do.

The second claim is one that I think is implied by the previous quote, which is that we should favor more evidence, all else being equal. So, if the expected value of two interventions is similar, we should generally favor investing in interventions that have more evidence supporting them.

In a case involving Drug A and Drug B and Drug C, maybe we would say something like, "These are relevantly similar." In a case where you have a lot of evidence that Drug A and Drug B have kind of the effects that you saw, this might actually favor giving a more well known drug over a new one, that's only been shown in one trial to be effective.

I'm basically going to consider both of these claims, and whether expected value calculations can vindicate either or both of them. As you kind of know from the spoilers, I'm going to argue that it can support the first claim but it actually rejects the second. Okay. So, I want to turn to this notion of resilience, and how we represent how much evident you have, in terms of the credences you assign to propositions like "This drug will cure this disease."

Probabilities and resilience



- Case 1: The untested coin
- Case 2: The well tested coin
- ...until the coin has been flipped a million times

Figure 2

Take the first case, which is this untested coin. I've given you no information about how biased this coin is. It could be completely biased in favor of heads, it could be completely biased in favor of tails, or it could be a completely fair coin. You have no information to distinguish between any of these hypotheses. It seems like, in a case where you have no idea about what the bias of a coin is and I say to you, "What is the chance it lands heads on the next throw?" You're going to have to say, "It's about 50%," because you have no reason to favor a heads bias over a tails bias.

Now consider a difference case, which is the well-tested coin. The well tested coin, you flip it, you get the following sequence, "Heads, heads, heads, tails, heads, heads, tails, tails," until the coin has been flipped a million times. You had a very, very boring series of days with this coin.

In the first case, in answer to the question, "What's the probability that the coin will land heads in the next flip?" you should say, "0.5 or 50%." In the second case, where you tested the coin a bunch and it's come up heads roughly 50% of the time, tails roughly 50% of the time, you should also say that the next flip is 50% likely to be heads.

The difference in these cases is reflected in the resilience levels of your credences. One kind of simple formulation of resilience, I think we can get a bit more specific with this, but for the purposes of this talk it doesn't matter too much, is that credo-resilience is how stable you expect your credences to be in response to new evidence. If my credences are high resilience, then there's more stability. I don't expect them to vary that much as new evidence comes in, even if the evidence is good and pertinent to the question. If they're low resilience, then they have low stability. I expect them to change a little in response to new evidence. That's true in the case of the untested coin, where I just have no data about how good it is, so the resilience of my credence of 50% is fairly low.

It's worth noting that resilience levels can reflect either the set of evidence that you have about a proposition, or your prior about the proposition. So, if it's just incredibly plausible that the coins are generally fair. For example, if you saw me simply pick the coin up out of a stack of otherwise fair coins, in this case you would have evidence that it's fair. But if you simply live in a world that doesn't include a lot of very biased coins, then your prior might be doing a lot of the work that your evidence would otherwise do. These are the two things that generate credo-resilience.

In both cases, with the coin, your credence that the coin will land heads on the next flip is the same, it's 0.5. Your credence of 0.5 about the tested coin is resilient, because you've done a million trials of this coin. Whereas, your credence about the untested coin is quite fragile. It could easily move in response to

Probabilities and resilience

- Case 3: You start to test the untested coins

Figure 3

new evidence, as we see here.

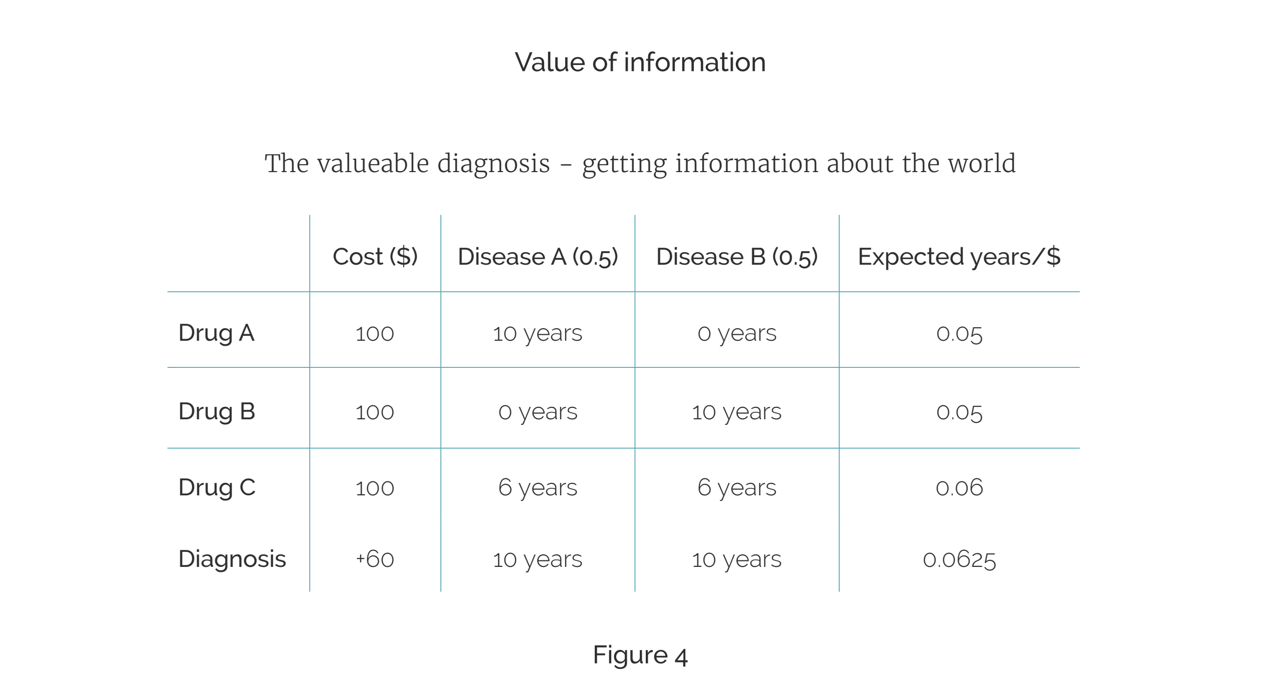
Take this third case. You start to test the untested coin, so you perform a series of flips with the coin, and you start to see a pattern [Figure 3]. In a case like this, it looks like the coin in front of you is pretty heavily heads biased, or you at least start to quite rapidly increase your credence that it's heads biased. So, your credence that it's going to come up heads next time is much higher. Because you had less evidence before, this credence was much more fragile, so now you've seen a change.

This would not happen if you got this sequence on the well-tested coin, because more evidence means that your credences are more resilient. If you saw a series of five head after performing a million trials, and it lands heads roughly half the time, this is just not going to make a huge difference to what you expect the next coin flip to be.

I think credo-resilience has some interesting effects. A lot of people seems to be kind of unwilling to assert probability estimates about whether something is going to work or not. I think a really good explanation for this is that, in cases where we don't have a lot of evidence, our credences about how good our credences are, are fairly low.

We basically think it's really likely that we're going to move around a lot in response to new evidence. We're just not willing to assert a credence that we think is just going to be false, or inaccurate once we gain a little bit more evidence. Sometimes people think you have mushy credences, that you don't actually have precise probabilities that you can assign to claims like, "This intervention is effective to Degree N." I actually think resilience might be a good way of explaining that away, to say, "No. You can have really precise estimates. You just aren't willing to assert them."

One thing that this has a huge influence on, and is kind of the theme to this talk, is the value of information. To return to our drug case, which I hope you'll see, the idea is that this is supposed to be somewhat analogous to interventions. However I don't want to put any interventions there, because I



Value of information

The valueable diagnosis – getting information about the world

	Cost ($)	Disease A (0.5)	Disease B (0.5)	Expected years/$
Drug A	100	10 years	0 years	0.05
Drug B	100	0 years	10 years	0.05
Drug C	100	6 years	6 years	0.06
Diagnosis	+60	10 years	10 years	0.0625

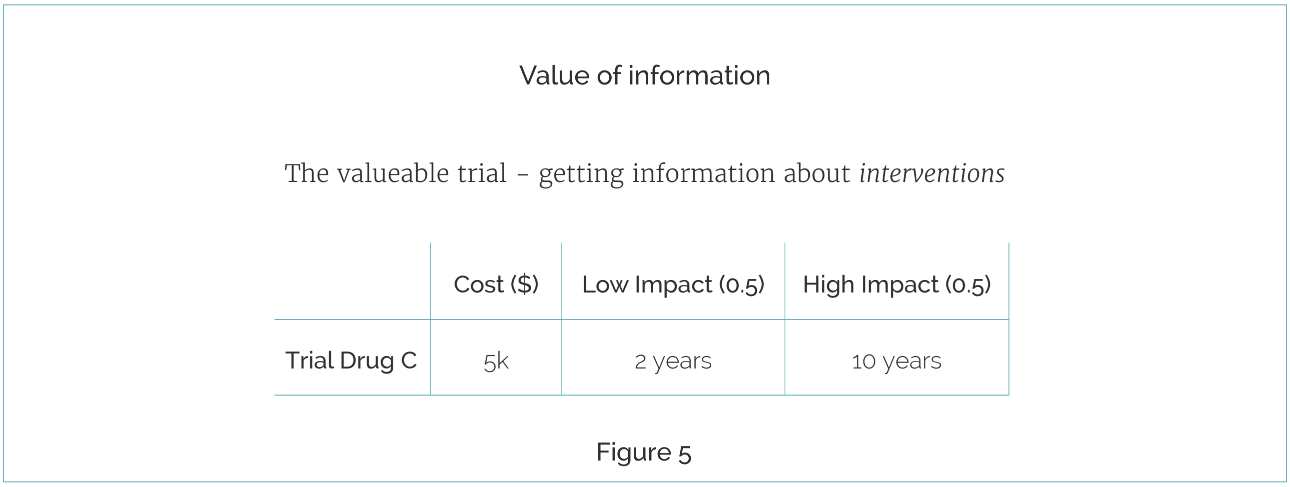
Figure 4

don’t want to make people think that I think their interventions don’t have enough evidence behind them.

In the original case, we had the following kind of scenario, where we had expected 0.05, 0.05 and 0.06 for the three drugs. Of course, one thing that we can do here is gain valuable evidence about the world. Consider this case, where diagnosis is invented, at least as far as Disease A and Disease B are concerned. So, we can now diagnose whether you have Disease A or Disease B, and it costs 60 additional dollars to do so. Given this, if I diagnose you, then I can expect that conditional on diagnosis, if you have Disease A, you will live for 10 years because, I will be able to then pay an additional $100 to give you Drug A. If you have Disease B, I’ll be able to pay an additional $100 to get you Drug B.

So in this case, the value of diagnosis, including the cost of then curing you of the disease, is actually higher than any of the original interventions. Rather than giving you Drug A, Drug B, Drug C, I should diagnose you and give you the correct drug. Hopefully this is intuitive.
Okay. That was information about the world, which maybe we think is valuable anyway. Supposed I care about global poverty and I want to find out good interventions, I can find out about deficiencies that exist in India for example. Then, I can see if there are good ways to improve that. So, that’s finding out about the world.

Obviously a different way we can gain valuable information here is by finding out about interventions themselves [Figure 5]. An example would be to look at the actual intervention of Drug C, and at how effective it is.



Value of information

The valueable trial – getting information about *interventions*

	Cost ($)	Low Impact (0.5)	High Impact (0.5)
Trial Drug C	5k	2 years	10 years

Figure 5

Suppose that the cost of a trial of Drug C, that would basically bring you to certainty about its effectiveness, is in this wonderfully ideal world, $5,000. Somehow, you know that it can only be very low impact or very high impact. You have a credence of about 0.5, that it’s going to come out that, in both Disease A and Disease B, this only actually extends life by two years. Let that be the kind of skeptical prior. But you also have the credence of about 0.5 that will extend life by 10 years in both cases. Let’s assume diagnosis has gone out the window.

Okay. So, you’re currently prescribing Drug C. You’re ignoring the fact that there’s low evidence here. You obviously don’t exist in any modern medical system. You’re going with the expected value as is, basically. Then the question is, “What is the value of doing this trial, especially given that you’re already prescribing Drug C?” If your credence of low impact goes to one, i.e., you suddenly discover that this drug is much

Value of information

The valueable trial – getting information about *interventions*

	Cost ($)	Low Impact is true (0.5)	High Impact is true (0.5)
No Trial	0	2 years/ $100	10 years/ $100
Trial	5k	5 years/ $100	10 years/ $100

Figure 6

less effective than you thought it would be, then you're going to switch from Drug C to prescribing Drug A or B again.

The per patient benefit is going to go from two years of expected life to five years of expected life in this case. Whereas, if you perform no trial, you won't spend anything, but you'll only get two years of additional life per $100, if it is in fact a low impact drug. You'll be prescribing Drug C continuously every time you see Disease A and Disease B, and it'll only give people an additional two years of life. Whereas if it's high impact, then you'll still get the ten year benefit because you'll have been accidentally prescribing something that's very good.

So, the trial adds 1.5 years of expected life per future treatment. The trial is therefore better than just prescribing Drug C if there are more than 2,000 patients. So, if there are more than 2,000 patients, the value of investing in a trial of Drug C is better than giving any of the drugs currently present. That means that the information value actually swamps the direct value of intervention here.

The value of investing in Drug A or B testing is going to be negligible because credences about their effectiveness are already resilient. So, this all builds up to what I think is the result of this, which is a really intrusive one really, which is that actually expected utility theory or expected value calculations might say that in cases where all else is equal, we should favor investing in interventions that have less evidence, rather than interventions that have more.

That means, if the expected concrete value of two interventions is similar, we should generally favor investing in interventions that have less evidence supporting them. I'm going to use concrete value to just mean "Non-informational value." The idea here is that in such cases, the concrete value is the same but, the information value for one of them is much higher, namely the one where you have a much lower resilience credence generating your expected value calculation.

So, this gets us to the "What does this mean, and what should we do" part. Hopefully I've convinced you that, despite the fact that it was an intuitive proposition, that we should favor things with more evidence, there's actually some argument that we should favor things that have less evidence.

Explore, Exploit or Evade

When we want to do good we can:

Explore: Invest resources in intervention x for its information value

- research, funding to gather data, career trials

Exploit: Invest resources in intervention x for its concrete value

- large project grants, career choices

Evade: Do not invest in intervention x

- investing elsewhere, delaying investment

Figure 7

When we're considering information value, there are basically three options available to us. I used to call this "Look, leap and retreat," and then I discovered that I really like things that sound the same so, I went for "Explore, exploit or evade" [Figure 7].

So we can choose to explore, and this means investing resources in interventions primarily for their information value. So things like research, funding to gather data, career trials. We can exploit, which means investing resources in an intervention for its concrete value. That means, things like large project grants and entire career choices. Or we can evade. We can decide just not to invest in a given intervention. We either invest elsewhere, or we completely delay investment.

The main difference between these is the reason for action. Take three people. Amy donates $100 to an existential risk charity to protect the future of humanity, so she's just exploiting the value. She's just looking at the direct concrete value of this intervention.

Bella donates $100 to the same charity to find out how much good they can do in the world. So, she's doing it mainly to explore, and then later she'll exploit. She'll think to herself "Okay. Let's see how valuable this is." If it's very valuable, then she'll basically mine the value of it.

Carla donates $100 to Charity C, so that we have more time to discover what the best causes are. So she is exploiting currently, by investing in it for its direct value for reducing existential risk, so that we can find out what the best cause is, so that we can exploit that. So, she's exploiting to explore to exploit.

So, when is exploring especially cost effective? Essentially, when there are three features. When there's more uncertainty about the direct value of an intervention, so this means options that have high expected value, but low resilience. When there are high benefits of certainty about the direct value, so when we can basically repeatedly mine something for value. And when there are low information costs, so when information's not too costly to obtain and the delay is low cost (you don't really want to be looking for information when cars are driving towards you, as the cost of not just taking action and getting out of the way is pretty high!).

The question I have is basically "Is gaining information especially valuable for effective altruists?" Maybe a different way to put this is "Is information essentially its own cause area, within effective altruism?"

There's a lot of uncertainty within and across good cause areas, especially if we consider long-term indirect effects. We don't know about the long-term indirect effects of a lot of our interventions. So, I think there's a lot of uncertainty here. We see that in terms of the progress that EA Research has made. I think this is evidence of a lot of uncertainty.

There are also high benefits of certainty in this case. We expect to use this information in the long term. Effective altruism isn't like a short-term intervention. So, in multi-generational projects, you expect the value of information to be higher, because people can essentially explore for longer and find optimal interventions.

To some degree, there are low information costs, so far as the movement is young, and there's still a lot of low-hanging fruit. This is with caveats: Maybe you're a bit like Carla. Maybe you're very worried that we're just screwing up the climate or that nuclear war is going to go terribly wrong. In which case, maybe you think we should just be directly intervening in those areas.

So, what difference would exploring more make to effective altruism, if you can abide my argument that this is an important cause area? Well, I think we could probably invest a lot more time and resources in interventions that are plausibly good, in order to get more evidence about them. I think we should probably do more research, but I realize that this point is kind of self-serving. I think that larger donors should probably diversify their giving more, if the value of information diminishes steeply enough, which I think might be the case.

Psychologically, I think we should be a bit more resilient to failure and change. I think when people consider the idea that they might be giving to cause areas that could turn out to just be completely fruitless, they find it psychologically difficult. In some ways just thinking "Look, I'm just exploring this to get the information about how good it is, and if it's bad, I'll just change. Or, if it doesn't do as well as I thought, I'll just change." I actually find this quite psychologically comforting, if you worry about these things.

The extreme view that you could have is "We should just start investing time and money in interventions with high expected value, but little or no evidential support." A more modest proposal, which is the one that I'm going to kind of endorse, is "We should probably start explicitly including the value of information, and assessments of causes and interventions, rather than treating it as an afterthought to concrete value." With some of things that I've looked at, I really think information value can swamp concrete value. If that's the case, it really shouldn't be an afterthought. It should be one of the primary drivers of values, not an afterthought in your calculation summary.

In summary, evidence does make a difference to expected value calculations via the value of information. If the expected concrete value to interventions is the same, this will favor testing out the intervention with less evidential support, rather than one with more. And taking value of information seriously would change what effective altruists invest their resources, i.e., their time and money in.

Question: One person did have a question about what it means to have credence in a credence. Maybe 80% chance that it has 50% chance of it working, etc., etc. Does it recourse down to zero, was the person's question.
Amanda Askell: It's not that you have a credence that you have a credence, but your credence in your credence being the same or changing in response to new evidence. There are a lot of related concepts here. There are things like "Your credence about the accuracy of your credence." So, it's not "I have a credence that I have a credence of 0.8." This is a separate thing, my credence that in response to this trial, I will adjust my credence from 0.5 to either 0.7 or 0.2, is the kind of credence that I'm talking about.

Question: Do you think there's a way to avoid falling into the rabbit hole, of the nesting credences of the kind that the person might have been referring to?

Amanda Askell: I guess my view, in the boring philosophical jargon, is that credences are dispositional. So, I do think that you probably have credences over infinitely many propositions. I mean, if I actually ask you about the proposition, you'll give me an answer. So, this is a really boring kind of answer, which is to say "No, the rabbit hole totally exists and I just try and get away from it by giving you a weird non-psychological account of credences."

Question: I'll take the side step. I'm not sure I'm parsing the question correctly but, I'll give it a go. They say, "Is information about the resilience captured by a full description of your current credences across the hypothesis space? If not, is there a parsimonious way to convey the extra information about resilience?"

Amanda Askell: Okay. I'm trying to think about the best way of parsing that. So, your credences across the hypothesis in question. Let's imagine that I'm just asking your credence, I say that the intervention has value N, for each N I'm considering. That will not capture the resilience of your credence because, it's going to be how you think that's going to adjust in response to a new state. If you include that how things are going to adjust in response to a new state in your hypotheses space, then yes, that should cover resilience. So yeah, it just depends on how you're carving up the space.

The Long-Term Future

Effective altruism often emphasizes the long-term impact of different interventions.

These articles explain why.

The Long-Term Future

By Jess Whittlestone, 2017

Introduction

The number of people alive today pales in comparison to the number who could exist in the future. It may therefore be extremely important to ensure that human civilisation flourishes far into the future.

There are a number of ways we might work to ensure a positive future for humanity. We could work to better understand and prevent extinction risks - catastrophic events that have the potential to destroy all life on this planet.[1] We may want to focus on the broader category of existential risks- events that could dramatically and irreversibly curtail humanity's potential.[2] Or we might focus on increasing the chance that the lives of our descendants are positive in other ways: for example, improving democracy or the ability of institutions to make good decisions.

Attempts to shape the long-term future seem highly neglected relative to the problems we face today. There are fewer incentives to address longer-term problems, and they can also be harder for us to take seriously.

It is, of course, hard to be certain about the impact of our actions on the very long-term future. However, it does seem that there are things we can do - and given the vast scale we are talking about, these actions could therefore have an enormous impact in expectation.

This profile sets out why you might want to focus your altruistic efforts on the long-term future - and why you might not. You may be particularly inclined to focus on this if you think we face serious existential threats in the next century, and if you're comfortable accepting a reasonable amount of uncertainty about the impact you are having, especially in the short-term.

The case for the long-term future as a target of altruism

The case for focusing on the long-term future can be summarised as follows:

1. The long-term future has *enormous potential*: our descendants could live for billions or trillions of years, and have very high-quality lives;
2. It seems likely there are *things we can do today* that will affect the long-term future in non-negligible ways;
3. Possible ways of shaping the long-term future are currently *highly neglected* by individuals and society;

[1] E.g. catastrophic climate change, nuclear war, or threats from advanced technologies.

[2] Failing to reach technological maturity is also classed as threatening the future of humanity, even though it may not sound like a particularly awful scenario, because of the huge loss of potential - "a technologically mature civilization could (presumably) engage in large-scale space colonization… be able to modify and enhance human biology… could construct extremely powerful computational hardware and use it to create whole-brain emulations and entirely artificial types of sentient, superintelligent minds" (Bostrom, 2012) - meaning the permanent destruction of this potential could constitute an enormous loss.

4. Given points 1 to 3 above, actions aimed at shaping the long-term future seem to have *extremely high* expected value, higher than any actions aiming for more near-term benefits.

Below we discuss each part of this argument in more detail.

The long-term future has enormous potential

Civilisation could continue for a billion years, until the Earth becomes uninhabitable.[3] It's hard to say how likely this is, but it certainly seems plausible – and putting less than, say, a 1% chance on this possibility seems overconfident.[4] Even if civilisation only survives for another million years, that still amounts to another ~50,000 generations of people, i.e. trillions of future lives.[5]

If our descendants survive for long enough, then they are likely to advance in ways we cannot currently imagine – even someone living a few hundred years ago could not possibly have imagined the technological advances we've made today. It is possible they might even develop technology enabling them to reach and colonise planets outside our solar system, and survive well beyond a billion years.[6]

Let's say that if we survive until the end of the Earth's lifespan, there is a 1% chance of space colonisation. This would make the overall probability of survival beyond Earth 1 in 10,000 (1% chance of surviving to a billion years, multiplied by a 1% chance of surviving further given that). This sounds incredibly low, but suppose that space colonisation could allow our descendants to survive up to 100 trillion years.[7] This suggests we could have up to 1/10,000 x 100 trillion years = 10 billion expected years of civilisation ahead of us.

If we expect life in the future to be, on average, about as good as the present, then this would make the whole of the future about 100 million times more important than everything that has happened in the last 100 years. In fact, it seems like there could be more people in the future with better lives than those living today: economic, social, and technological progress could enable us to cure diseases, lift people out of poverty, and better solve other problems. It also seems possible that people in the future will be more altruistic than people alive today[8] – which also makes it more likely that they will be motivated to create a happy and valuable world.

However, it's precisely because of this enormous potential that it's so important to ensure that things go as well as possible. The loss of potential would be enormous if we end up on a negative trajectory. It could

[3] "It is not absurd to consider the possibility that civilization continues for a billion years, until the Earth becomes uninhabitable" Nick Beckstead, "On the Overwhelming Importance of Shaping the Far Future"

[4] Note that we don't necessarily care about just the species Homo Sapiens - when we talk about our "descendants", we mean any valuable successors we might have, and include in this non-human animals that seem to warrant moral concern.

[5] Assuming one generation every 20 years, and ~75 million people per generation.

[6] Nick Beckstead, reviewing expert opinion on the topic, concludes that, "most informed people thinking about these issues believe that space colonization will eventually be possible" https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/

[7] "Stars will continue shining for about 10^14 more years" Adams, 2008

[8] Though this is a claim we will not defend in detail here, it certainly seems like our "circle of compassion" has expanded over time, and Stephen Pinker presents a compelling case that violence is decreasing.

result in a great deal of suffering or the end of life.[9] And just as the potential to solve many of the world's problems is growing, threats seem to be growing too. In particular, advanced technologies and increasing interconnectedness pose great risks.[10]

There are things we can do today that could affect the long-term future

There are a number of things we could work on today that seem likely to influence the long-term future:

- *Reducing extinction risks*: We could reduce the risk of catastrophic climate change by putting in place laws and regulations to cut carbon emissions. We could reduce the risks from new technologies by investing in research to ensure their safety. Alternatively, we could work to improve global cooperation so that we are better able to deal with unforeseen risks that might arise.
- *Changing the values of a civilisation*: Values tend to be stable in societies, so attempts to shift values, whilst difficult, could have long-lasting effects. Some forms of value change, like increasing altruism, seem robustly good, and may be a way of improving the future. However, some forms of value spreading may be harmful.
- *"Speeding up" development*: Boosting technological innovation or scientific progress could have a lasting "speed up" effect on the entire future, making all future benefits happen slightly earlier than they otherwise would have. Curing a disease just a few years earlier could save millions of lives, for example. (That said, it's not clear whether speeding up development is good or bad for existential risk - developing new technologies faster might help us to mitigate certain threats, but pose new risks of their own.)
- *Ripple effects of our ordinary actions*: Improvements in health not only benefit individuals directly but allow them to be more economically successful, meaning that society and other individuals have to invest less in supporting them. In aggregate, this could easily have substantial knock-on effects on the productivity of society, which could affect the future.
- *Other ways we might create positive trajectory changes*: These include improving education, science, and political systems.

Paul Christiano also points out that even if opportunities to shape the long-term future with any degree of certainty do not exist today, they may well exist in the future. Investing in our own current capacity could have an indirect but large impact by improving our ability to take such opportunities when they do arise. Similarly, we can do research today to learn more about how we might be able to impact the long-term future.

The long-term future is neglected, especially relative to its importance

Attempts to shape the long-term future are neglected by individuals, organisations and governments.

One reason is that there is little incentive to focus on far-off, uncertain issues compared to more certain, immediate ones. As 80,000 Hours put it, "Future generations matter, but they can't vote, they can't buy

[9] It might seem that the risk of extinction is low. However, Nick Bostrom writes that "estimates of 10-20% total existential risk in this century are fairly typical among those who have examined the issue, though inevitably such estimates rely heavily on subjective judgement."

[10] The Open Philanthropy Project suggests that "as the world becomes more interconnected, the magnitude and implications of the worst-case scenarios may be rising."

things, they can't stand up for their interests."

Problems faced by future generations are also more uncertain and more abstract, making it harder for us to care about them. There is a well-established phenomenon called temporal discounting, which means that we tend to give less weight to outcomes that are far in the future. This may explain our tendency to neglect long-term risks and problems. For example, it's a large part of why we seem to have such difficult tackling climate change.

Generally, there are diminishing returns to additional work on an area. This means that the neglectedness of the long-term future makes it more likely to be high impact.

Efforts to shape the long-term future could be extremely high in expected value

Even if the chance of our actions influencing the long-term trajectory of humanity is relatively low, there are extremely large potential benefits, which mean that these actions could still have a very high expected value. For example, decreasing the probability of human extinction by just one in a million could result in an additional 1,000 to 10,000 expected years of civilisation (using earlier assumptions).[11]

Compare this to actions we could take to improve the lives of people alive today, without looking at longer-run effects. A dramatic victory such as curing the most common and deadly diseases, or ending all war, could probably only make the current time period (~100 years) up to twice as good as otherwise.[12] Though this seems like an enormous success, given the calculations above, decreasing the probability of human extinction would be 10 or 100 times better in expectation.

We might want to adjust this naive estimate downwards slightly, however, given uncertainty about some of the assumptions that go into it - we could be wrong about the probability of humanity surviving far into the future, or about the value of the future (if we think that future flourishing might have diminishing value, for example.) However, even if we think these estimates should be adjusted downwards *substantially*, we might very conservatively imagine that reducing the likelihood of existential risk by one in a million only equates to 100 expected years of civilization. This *still* suggests that the value of working to reduce existential risk is comparable to the value of the biggest victories we could imagine in the current time period - and so well worth taking seriously.

Some concerns about prioritising the long-term future

This is a counterintuitive way of doing good

Some people might be sceptical about this cause because focusing on the long-term future seems counterintuitive. Altruism normally means helping those around us - not trying to ensure that our descendants survive for billions of years. If the long-term future is so important, how come almost no-one is thinking or talking about it?

[11] It's worth noting here, however, that reducing the probability of extinction by one in a million may be harder than it sounds - given that there is not just one single threat that needs to be tackled, but many different scenarios that could seriously threaten human extinction.

[12] "A dramatic victory around the world might make this period go twice as well as it otherwise would, say"- Nick Beckstead, On the Overwhelming Importance of Shaping the Far Future

Firstly, just because something goes against common sense doesn't mean that it is mistaken. The moral intuitions of society have been wrong in the past – we now think of the way people used to treat slaves, women, and gay people as morally abhorrent, for example. There's also evidence that our moral intuitions are subject to a whole host of biases, such as being insensitive to differences when numbers are large – intuitively, helping 10,000 people feels similarly good to helping 1 million people, even though they are vastly different in scale.[13]

Secondly, it's not clear that prioritising the long-term future actually is that counter-intuitive, properly understood. The long-term perspective suggests that we should prioritise things like investing in technology and economic growth, scientific research, strengthening institutions and improving democracy. These seem like pretty plausible responses to the question, "How can we best solve the world's biggest problems?". In fact, to some they may seem more intuitive than focusing on more proximate benefits like giving cash to poor farmers in Kenya or distributing malaria nets.[14]

Shouldn't we prioritise already existing people?

Another objection to this cause might be that future, unborn people don't matter morally – at least not in the same way that people currently alive do.

It is easier to empathise with people who are already alive, and this might mean that we feel more motivated to help them. We might feel that it's easier to help current people, because we can learn what they need. However, none of this implies that future people actually matter any less, morally.

When pressed, it seems like most people do care about the wellbeing of future generations. 80,000 Hours gives the example of a simple choice between (a) preventing one person from suffering next year, and (b) preventing 100 people from suffering (the same amount) 100 years from now. Since most people would choose the second option, they suggest that this shows that most people do value future generations.

Is there anything we can do?

A final objection is that we can't influence the long-term future with any certainty, and so it's useless to even try. Our efforts would be better focused on problems where we can see progress more clearly.

In the last section we discussed reasons to think that we probably can influence the long-term future. It's worth reiterating that we don't need to be certain of the impact of our actions. Efforts to shape the future could impact billions of lives. So even only a 1% chance of success would seem worth taking a 'bet' on.

Of course, we do still need to show that there is a decent probability of success, or else we could justify any total long-shot idea with this reasoning. The point here is just that we don't need to be *highly confident*.

[13] Nick Beckstead has an extensive discussion of these reasons in chapter 2 of his thesis, "On the Overwhelming Importance of Shaping the Far Future" (available on his website here: http://www.nickbeckstead.com/research).

[14] Ben Todd makes this point in an article, entitled "Why long-run focused effective altruism is more common sense". He acknowledges that some very specific long-term future-oriented beliefs, like the idea that we should prioritise reducing existential risks from artificial intelligence, may be counter-intuitive. But simply focusing on the long-term perspective when doing good is common sensical.

Reasons you might not choose to prioritise the long-term future

You might question how much moral weight we should give to "future people"

We generally feel an intuitive obligation to treat future, not-yet-existent people in roughly the same way as existent people. The argument laid out above assumes that we assign the same "moral weight" to future people as those living today. However, some philosophers have questioned whether we should give the same degree of moral consideration to future people. On "person-affecting views", an action is only good or bad if it is good or bad for someone – and so the value of an action depends only on how it affects people who are either already alive, or who will come into existence regardless of our actions. This suggests that we have stronger reason to improve the lives of people living today than to help future generations. These views also suggest that human extinction, while bad for the people who die, causes no longer-term harms: there is no harm in people failing to come into existence.

A related issue is the non-identity problem, arising from the fact that sometimes future people may owe their very existence to choices made today. For example, which policies a government chooses to enact will affect which people have certain jobs, affecting which people meet and marry, and therefore causing different people to be born in future. Those very policies might also affect how good the lives of future people are – if the government chooses to prioritise policies that increase short-term economic productivity over mitigating climate change in the longer-term, say, this could have a negative impact on future generations. But if the very policies that appear to have made future people's lives "worse off" also ensured that those exact people were born at all, can those people really be said to have been harmed by those policies? If not, then this may be reason to prioritise the welfare of people who already exist, or whose existence does not depend on our actions.

However, the non-identity problem might also be taken as a reason to reject person-affecting views. The implication, that choosing policies that will make future generations lives worse off is not causing those future people any harm, seems highly counterintuitive. We could instead adopt impersonal principles for evaluating the moral value of actions.[15] This means that we would judge not based on how they affect specific people, but based on how good they are from the perspective of the world as a whole.[16]

You might think that we can't affect the long-term future

You might think that there is nothing we can do that has a reliable impact on the long-term future.[17]

For example, you may agree that our actions have indirect effects, but deny that we can tell what those effects will be in advance. It is much easier to look back and identify past actions which led to substantial

[15] This is roughly the perspective taken by philosopher Derek Parfit, who says that it's so obviously good for us to help future generations that this itself gives us a definitive reason to reject person-affecting principles.

[16] Another possibility, which Parfit discusses, is to adopt wider person-affecting principles: these say that while most harms / benefits are comparative (i.e. to say we harmed a person is to say they would have been better off had we acted differently), not all are. In particular, we can say that someone was benefitted by being brought into existence if their life is overall positive. On such views, one state of the world is worse than another, even if different people exist in the two worlds (and therefore neither state is strictly speaking worse for anyone), if the lives of the people in World A are less good for those people than the lives of people in World B are for them.

[17] Note that this is subtly different from saying that our actions today don't affect anything in the future (which would seem hard to argue for). Instead, it is the claim that we cannot reliably or usefully estimate the impact of our actions on the future.

knock-on benefits than it is to predict what actions will have those effects in future. Or you may think that we cannot make meaningful predictions about existential risks. Or you may think that there is little we can do about existential risks given our current knowledge, institutions, and technology.

Though it is very difficult to be confident here, there are opportunities which seem likely to have some impact. The fact that we are uncertain means that it may also be worth investing in our ability to identify more valuable opportunities to improve the future.

You might disagree with "expected value" reasoning, in principle or in practice

The final step in the argument laid out above relies on the notion of the expected value of an action. The expected value of an action combines (a) the value of each possible outcome with (b) the probability of each outcome occurring. This allows us to see why it can sometimes be better to take an action that has a smaller chance of success, but a greater reward if successful.

It's worth separating out two subtly different ways that you might disagree with the use of expected values in this argument.

1. Objecting to the use of expected values to make decisions in principle

Some people challenge the use of expected value at all. Expected value theory can run into problems especially if we allow the possibility of infinite amounts of value (since expected value theory says that any non-zero probability of creating infinite value should dominate our decision-making).

The main alternative to expected value theory is pure risk aversion.[18] This view states that we should sometimes take lower expected value options if they offer extra certainty.[19] Risk aversion is a matter of degree: it doesn't specify exactly what tradeoff we should make between certainty and potential impact. Introducing any amount of risk aversion weakens the case for the long-term future.

Should we be risk averse when we evaluate altruistic actions? Risk aversion seems to be irrational in a fairly fundamental sense - that is, risk averse agents can end up having inconsistent preferences that leave them open to exploitation[20] (though note that Buchak does argue that risk aversion can be rational).[21]

2. Objecting to the way that expected values are estimated in practice

You might agree that expected values are how we should make decisions under uncertainty in *principle*, but disagree with the way that these expected values are calculated in *practice*.

There are always difficulties in calculating expected value. But the difficulties expand when we lack historical data or studies on the intervention. It therefore seems reasonable to be sceptical of any attempt

[18] Though note that you need an extreme form of risk aversion to avoid the problems with infinite value mentioned above.

[19] Note that by pure risk aversion we mean that people are risk averse over the unit of value. This is different from monetary risk aversion, which arises from the fact that there are generally diminishing marginal returns to the utility of income.

[20] De Finetti, B. (1964). Foresight: its logical laws in subjective sources; Ramsey, F. P. (1926). Truth and probability. The foundations of mathematics and other logical essays, pages 156–198.

[21] Buchak, L. (2009). Risk Aversion and Rationality.

to argue from the expected value of an intervention in this space.

Often when we're reasoning, it makes sense to account for base rates. In this case, the base rate is the average effectiveness of a cause area. Suppose you think that the average cause area effectiveness is much lower than our estimate of the effectiveness of work to improve the future. Since our estimate is quite speculative, it seems like we should think it's pretty possible that we've made a mistake, and this means that our all-things considered judgement about the effectiveness of work on the long-term future should be lower than our initial calculation suggests.[22]

We think that the case we've made for prioritising the long-term future is based on wider considerations than simple expected value estimates (e.g. we've argued that it makes intuitive sense, and discussed a number of different heuristics which point in this direction). However you might certainly put less weight on any *specific* attempts to shape the long-term future if you are sceptical of our ability to estimate EV in this area.

You might think that solving current problems is the best way to influence the long-term future

Finally, you might agree with the case laid out above, but believe that the best way to improve the future is by solving the biggest problems that exist today. This isn't, strictly speaking, an objection to the general idea that the long-term future of humanity should be a priority. It's just a view about the *best way* to improve the future. As we've discussed, there's reason to think that solving problems today will have hard-to-anticipate long-term effects. For instance, reducing poverty will improve health and wellbeing. But it may also have significant long-term effects, in that more people will be able to contribute productively to society, to innovation, to the economy, etc.

You might also choose to focus on current problems if you think that these problems are unlikely to be solved by general progress. One worry is that if we leave current problems unsolved, we might get locked into bad structures. For instance, it might be important to solve inequality, in case the problem gets harder to solve later.[23]

You might therefore choose to focus on solving more immediate, concrete problems in the world even if you think that the long-term is ultimately what's most important, if you believe (most of) the following:

1. There is no immediate risk of extinction, or there is nothing much we (or you personally) can do about it;
2. Solving the biggest problems we currently face as a society is likely to have large, long-term benefits, and there's nothing else we can do which would have larger long-term benefits;
3. These problems may not get solved indirectly if we focus on becoming more powerful and

[22] This is called Bayesian updating. We should be similarly sceptical of cost-effectiveness estimates in other areas, like global health, but this will be less important for such areas because the evidence is more robust.

[23] Another risk would be humans becoming more and more powerful without a similar rate of increase in empathy towards less powerful species - this could result in our descendants causing more suffering for other animals even if only accidentally. We probably cause more animal suffering today than we did a few hundred years ago, for example, not because we care less about animals (we probably care more), but because it is much easier for us to farm animals en masse in conditions with little regard for their welfare.

knowledgeable as a species, and may even worsen.

Summary

- The long-term future has enormous potential: our ancestors could live for billions or trillions of years, and their lives could be very good.
- It seems likely that there are things we can do today that will affect the long-term future in non-negligible ways.
- Attempts to shape the long-term future are currently highly neglected by individuals and society.
- Given the above, actions aimed at shaping the long-term future seem to have extremely high expected value, higher than any actions aiming for more near-term benefits.
- This argument relies on the assumption that we should value future people in roughly the same way we value people currently alive – an assumption which is challenged by *person-affecting views* of ethics.
- It also relies heavily on "expected value" calculations, which might be questioned either in principle or practice – if you think we should be *risk averse* about value, or if you think it's likely that a mistake may have gone into the expected value estimate in this case.
- It's also possible that solving current problems is the best way to affect the far future – if you think there's little or no risk of human extinction, and progress on more immediate problems is likely to have large long-term benefits.

A Proposed Adjustment to the Astronomical Waste Argument

By Nick Beckstead, 2013

This article was originally published as A Proposed Adjustment to the Astronomical Waste Argument on lesswrong.com. In this 2013 article, Nick Beckstead argues that Nick Bostrom's argument in his paper "Astronomical Waste: The Opportunity Cost of Delayed Technological Development" should be adjusted. Beckstead argues that a focus on the long-term future does not entail targeted and narrow work to reduce existential risk, but that we instead also should consider broad interventions to improve the trajectory to the future.

An existential risk is a risk "that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development," (Bostrom, 2013). Nick Bostrom has argued that

"The loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole. It may be useful to adopt the following rule of thumb for such impersonal moral action:

Maxipok: Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe."

There are a number of people in the effective altruism community who accept this view and cite Bostrom's argument as their primary justification. Many of these people also believe that the best ways of minimizing existential risk involve making plans to prevent specific existential catastrophes from occurring, and believe that the best giving opportunities must be with charities that primarily focus on reducing existential risk. They also appeal to Bostrom's argument to support their views. (Edited to add: Note that Bostrom himself sees maxipok as neutral on the question of whether the best methods of reducing existential risk are very broad and general, or highly targeted and specific.) For one example of this, see Luke Muehlhauser's comment:

"Many humans living today value both current and future people enough that if existential catastrophe is plausible this century, then upon reflection (e.g. after counteracting their unconscious, default scope insensitivity) they would conclude that reducing the risk of existential catastrophe is the most valuable thing they can do — whether through direct work or by donating to support direct work."

I now think these views require some significant adjustments and qualifications, and given these adjustments and qualifications, their practical implications become very uncertain. I still believe that what matters most about what we do is how our actions affect humanity's long-term future potential, and I still believe that targeted existential risk reduction and research is a promising cause, but it now seems unclear whether targeted existential risk reduction is the best area to look for ways of making the distant future go as well as possible. It may be and it may not be, and which is right probably depends on many messy details about specific opportunities, as well as general methodological considerations which are, at this point, highly uncertain. Various considerations played a role in my reasoning about this, and I intend to talk about more of them in greater detail in the future. I'll talk about just a couple of these considerations in this post.

In this post, I argue that:

1. *Though Bostrom's argument supports the conclusion that maximizing humanity's long term potential is extremely important, it does not provide strong evidence that reducing existential risk is the best way of maximizing humanity's future potential.* There is a much broader class of actions which may affect humanity's long-term potential, and Bostrom's argument does not uniquely favor existential risk over other members in this class.
2. *A version of Bostrom's argument better supports a more general view: what matters most is that we make path-dependent aspects of the far future go as well as possible.* There are important questions about whether we should accept this more general view and what its practical significance is, but this more general view seems to be a strict improvement on the view that minimizing existential risk is what matters most.
3. *The above points favor very broad, general, and indirect approaches to shaping the far future for the better*, rather than thinking about very specific risks and responses, though there are many relevant considerations and the issue is far from settled.

I think some prominent advocates of existential risk reduction already agree with these general points, and believe that other arguments, or other arguments together with Bostrom's argument, establish that direct existential risk reduction is what matters most. This post is most relevant to people who currently think Bostrom's arguments may settle the issues discussed above.

Path-dependence and trajectory changes

In thinking about how we might affect the far future, I've found it useful to use the concept of the world's *development trajectory*, or just trajectory for short. The world's development trajectory, as I use the term, is a rough summary way the future will unfold over time. The summary includes various facts about the world that matter from a macro perspective, such as how rich people are, what technologies are available, how happy people are, how developed our science and culture is along various dimensions, and how well things are going all-things-considered at different points of time. It may help to think of the trajectory as a collection of graphs, where each graph in the collection has time on the x-axis and one of these other variables on the y-axis.

With that concept in place, consider three different types of benefits from doing good. First, doing something good might have *proximate benefits*—this is the name I give to the fairly short-run, fairly predictable benefits that we ordinarily think about when we cure some child's blindness, save a life, or help an old lady cross the street. Second, there are benefits from *speeding up development*. In many cases, ripple effects from good ordinary actions speed up development. For example, saving some child's life might cause his country's economy to develop very slightly more quickly, or make certain technological or cultural innovations arrive more quickly. Third, our actions may slightly or significantly alter the world's development trajectory. I call these shifts *trajectory changes*. If we ever prevent an existential catastrophe, that would be an extreme example of a trajectory change. There may also be smaller trajectory changes. For example, if some species of dolphins that we really loved were destroyed, that would be a much smaller trajectory change.

The concept of a trajectory change is closely related to the concept of path dependence in the social sciences, though when I talk about trajectory changes I am interested in effects that persist much longer than standard examples of path dependence. A classic example of path dependence is our use of QWERTY keyboards. Our keyboards could have been arranged in any number of other possible ways. A large part of the explanation of why we use QWERTY keyboards is that it happened to be convenient for making typewriters, that a lot of people learned to use these keyboards, and there are advantages to having most people use the same kind of keyboard. In essence, there is path dependence whenever some aspect of the future could easily have been way X, but it is arranged in way Y due to something that happened in the past, and now it would be hard or impossible to switch to way X. Path dependence is especially interesting when way X would have been better than way Y. Some political scientists have argued that path dependence is very common in politics. For example, in an influential paper (with over 3000 citations) Pierson (2000, p. 251) argues that:

"Specific patterns of timing and sequence matter; a wide range of social outcomes may be possible; large consequences may result from relatively small or contingent events; particular courses of action, once introduced, can be almost impossible to reverse; and consequently, political development is punctuated by critical moments or junctures that shape the basic contours of social life."

The concept of a trajectory change is also closely related to the concept of a historical contingency. If Thomas Edison had not invented the light bulb, someone else would have done it later. In this sense, it is not historically contingent that we have light bulbs, and the most obvious benefits from Thomas Edison inventing the light bulb are proximate benefits and benefits from speeding up development. Something analogous is probably true of many other technological innovations such as computers, candles, wheelbarrows, object-oriented programming, and the printing press. Some important examples of historical contingencies: the rise of Christianity, the creation of the US Constitution, and the writings of Karl Marx. Various aspects of Christian morality influence the world today in significant ways, but the fact that those aspects of morality, in exactly those ways, were part of a dominant world religion was historically contingent. And therefore events like Jesus's death and Paul writing his epistles are examples of trajectory changes. Likewise, the US Constitution was the product of deliberation among a specific set of men, the document affects government policy today and will affect it for the foreseeable future, but it could easily have been a different document. And now that the document exists in its specific legal and historical context, it is challenging to make changes to it, so the change is somewhat self-reinforcing.

Some small trajectory changes could be suboptimal

Persistent trajectory changes that do not involve existential catastrophes could have great significance for shaping the far future. It is unlikely that the far future will inherit many of our institutions exactly as they are, but *various aspects of the far future—including social norms, values, political systems, and perhaps even some technologies—may be path dependent on what happens now, and sometimes in suboptimal ways.* In general, it is reasonable to assume that if there is some problem that might exist in the future and we can do something to fix it now, future people would also be able to solve that problem. But if values or social norms change, they might not agree that some things we think are problems really are problems. Or, if people make the wrong decisions now, certain standards or conventions may get entrenched, and resulting problems may be too expensive to be worth fixing. For further categories of examples of path-dependent aspects of the far future, see these posts by Robin Hanson.

The astronomical waste argument and trajectory changes

Bostrom's argument only works if reducing existential risk is the most effective way of maximizing humanity's future potential. But *there is no robust argument that trying to reduce existential risk is a more effective way of shaping the far future than trying to create other positive trajectory changes.* Bostrom's argument for the overwhelming importance of reducing existential risk can be summarized as follows:

1. The expected size of humanity's future influence is astronomically great.
2. If the expected size of humanity's future influence is astronomically great, then the expected value of the future is astronomically great.
3. If the expected value of the future is astronomically great, then what matters most is that we maximize humanity's long-term potential.
4. Some of our actions are expected to reduce existential risk in not-ridiculously-small ways.
5. If what matters most is that we maximize humanity's future potential and some of our actions are expected to reduce existential risk in not-ridiculously-small ways, what it is best to do is primarily determined by how our actions are expected to reduce existential risk.
6. Therefore, what it is best to do is primarily determined by how our actions are expected to reduce existential risk.

Call that the "astronomical waste" argument.

It is unclear whether premise (5) is true because it is unclear whether trying to reduce existential risk is the most effective way of maximizing humanity's future potential. For all we know, it could be more effective to try to create other positive trajectory changes. Clearly, it would be better to prevent extinction than to improve our social norms in a way that indirectly makes the future go one millionth better, but, in general, "X is a bigger problem than Y" is only a weak argument that "trying to address X is more important than trying to address Y." To be strong, the argument must be supplemented by looking at many other considerations related to X and Y, such as how much effort is going into solving X and Y, how tractable X and Y are, how much X and Y could use additional resources, and whether there are subsets of X or Y that are especially strong in terms of these considerations.

Bostrom does have arguments that speeding up development and providing proximate benefits are not as important, in themselves, as reducing existential risk. And these arguments, I believe, have some plausibility. *Since we don't have an argument that reducing existential risk is better than trying to create other positive trajectory changes and an existential catastrophe is one type of trajectory change, it seems more reasonable for defenders of the astronomical waste argument to focus on trajectory changes in general*. It would be better to replace the last two steps of the above argument with:

4'. Some of our actions are expected to change our development trajectory in not-ridiculously-small ways.

5'. If what matters most is that we maximize humanity's future potential and some of our actions are expected to change our development trajectory in not-ridiculously-small ways, what it is best to do is primarily determined by how our actions are expected to change our development trajectory.

6'. Therefore, what it is best to do is primarily determined by how our actions are expected to change our development trajectory.

This seems to be a strictly more plausible claim than the original one, though it is less focused.

In response to the arguments in this post, which I e-mailed him in advance, Bostrom wrote a reply (see the end of the post). The key comment, from my perspective, is:

"Many trajectory changes are already encompassed within the notion of an existential catastrophe. Becoming permanently locked into some radically suboptimal state is an xrisk. The notion is more useful to the extent that likely scenarios fall relatively sharply into two distinct categories---very good ones and very bad ones. To the extent that there is a wide range of scenarios that are roughly equally plausible and that vary continuously in the degree to which the trajectory is good, the existential risk concept will be a less useful tool for thinking about our choices. One would then have to resort to a more complicated calculation. However, extinction is quite dichotomous, and there is also a thought that many sufficiently good future civilizations would over time asymptote to the optimal track."

I agree that a key question here is whether there is a very large range of plausible equilibria for advanced civilizations, or whether civilizations that manage to survive long enough naturally converge on something close to the best possible outcome. The more confidence one has in the second possibility, the more interesting existential risk is as a concept. The less confidence one has in the second possibility, the more interesting trajectory changes in general are. However, I would emphasize that unless we can be highly confident in the second possibility, it seems that we cannot be confident that reducing existential risk is more important than creating other positive trajectory changes because of the astronomical waste argument alone. This would turn on further considerations of the sort I described above.

Broad and narrow strategies for shaping the far future

Both the astronomical waste argument and the fixed up version of that argument conclude that what matters most is how our actions affect the far future. I am very sympathetic to this viewpoint, abstractly considered, but I think its practical implications are highly uncertain. There is a spectrum of strategies for shaping the far future that ranges from the very targeted (e.g., stop that asteroid from hitting the Earth) to very broad (e.g., create economic growth, help the poor, provide education programs for talented youth), with options like "tell powerful people about the importance of shaping the far future" in between. The limiting case of breadth might be just optimizing for proximate benefits or for speeding up development. Defenders of the astronomical waste argument tend to be on the highly targeted end of this spectrum. I think it's a very interesting question where on this spectrum we should prefer to be, other things being equal, and it's a topic I plan to return to in the future.

The arguments I've offered above favor broader strategies for shaping the far future, though they don't settle the issue. The main reason I say this is that *the best ways of creating positive trajectory changes may be very broad and general, whereas the best ways of reducing existential risk may be more narrow and specific.* For example, it may be reasonable to try to assess, in detail, questions like, "What are the largest specific existential risks?" and, "What are the most effective ways of reducing those specific risks?" In contrast, it seems less promising to try to make specific guesses about how we might create smaller positive trajectory changes because there are so many possibilities and many trajectory changes do not have significance that is predictable in advance. No one could have predicted the persistent ripple effects that Jesus's life had, for example. In other cases—such as the framing of the US Constitution—it's clear that a decision has trajectory change potential, but it would be hard to specify, in advance, which concrete measures should be taken. In general, it seems that the worse you are at predicting some phenomenon that is critical to your plans, the less your plans should depend on specific predictions about that phenomenon. Because of this, promising ways to create positive trajectory changes in the world may be more broad than the most promising ways of trying to reduce existential risk specifically. Improving education, improving parenting, improving science, improving our political system, spreading humanitarian values, or otherwise improving our collective wisdom as stewards of the future could, I believe, create many small, unpredictable positive trajectory changes.

I do not mean to suggest that broad approaches are necessarily best, only that people interested in shaping the far future should take them more seriously than they currently do. The way I see the trade-off between highly targeted strategies and highly broad strategies is as follows. Highly targeted strategies for shaping the far future often depend on highly speculative plans, often with many steps, which are hard to execute. We often have very little sense of whether we are making valuable progress on AI risk research or geo-engineering research. On the other hand, highly broad strategies must rely on implicit assumptions about the ripple effects of doing good in more ordinary ways. It is very subtle and speculative to say how ordinary actions are related to positive trajectory changes, and estimating magnitudes seems extremely challenging. Considering these trade-offs in specific cases seems like a promising area for additional research.

Summary

In this post, I argued that:

1. The astronomical waste argument becomes strictly more plausible if we replace the idea of minimizing existential risk with the idea of creating positive trajectory changes.
2. There are many ways in which our actions could unpredictably affect our general development trajectory, and therefore many ways in which our actions could shape the far future for the better. This is one reason to favor broad strategies for shaping the far future.

The trajectory change perspective may have other strategic implications for people who are concerned about maximizing humanity's long-term potential. I plan to write about these implications in the future.[24]

[24] I am grateful to Nick Bostrom, Paul Christiano, Luke Muehlhauser, Vipul Naik, Carl Shulman, and Jonah Sinick for feedback on earlier drafts of this post.

Comment from Nick Bostrom on this post

What follows is an e-mail response from Nick Bostrom. He suggested to share his comment along with the post. Note that the author added a couple of small clarifications to the above post in response to Bostrom's comment.

One can arrive at a more probably correct principle by weakening, eventually arriving at something like 'do what is best' or 'maximize expected good'. There the well-trained analytic philosopher could rest, having achieved perfect sterility. Of course, to get something fruitful, one has to look at the world not just at our concepts.

Many trajectory changes are already encompassed within the notion of an existential catastrophe. Becoming permanently locked into some radically suboptimal state is an xrisk. The notion is more useful to the extent that likely scenarios fall relatively sharply into two distinct categories---very good ones and very bad ones. To the extent that there is a wide range of scenarios that are roughly equally plausible and that vary continuously in the degree to which the trajectory is good, the existential risk concept will be a less useful tool for thinking about our choices. One would then have to resort to a more complicated calculation. However, extinction is quite dichotomous, and there is also a thought that many sufficiently good future civilizations would over time asymptote to the optimal track.

In a more extended and careful analysis there are good reasons to consider second-order effects that are not captured by the simple concept of existential risk. Reducing the probability of negative-value outcomes is obviously important, and some parameters such as global values and coordination may admit of more-or-less continuous variation in a certain class of scenarios and might affect the value of the long-term outcome in correspondingly continuous ways. (The degree to which these complications loom large also depends on some unsettled issues in axiology; so in an all-things-considered assessment, the proper handling of normative uncertainty becomes important. In fact, creating a future civilization that can be entrusted to resolve normative uncertainty well wherever an epistemic resolution is possible, and to find widely acceptable and mutually beneficial compromises to the extent such resolution is not possible---this seems to me like a promising convergence point for action.)

It is not part of the xrisk concept or the maxipok principle that we ought to adopt some maximally direct and concrete method of reducing existential risk (such as asteroid defense): whether one best reduces xrisk through direct or indirect means is an altogether separate question.

Promising Causes

There are many promising areas to work on, and it is difficult to work out which is most promising.

These articles discuss the latest thinking in some of the most promising cause areas, and the reasons for and against working on them.

Three Impacts of Machine Intelligence

By Paul Christiano, 2013

In this article, Paul Christiano argues that the development of intelligent machines will lead to accelarated growth, to reduced human wages, and to the decisions of intelligent machines playing a part in shaping the future.

I think that the development of human level AI in my lifetime is quite plausible; I would give it more than a 1-in-4 chance. In this post I want to briefly discuss what I see as the most important impacts of AI. I think these impacts are the heavy hitters by a solid margin; each of them seems like a big deal, and I think there is a big gap to #4.

1. Growth will accelerate, probably very significantly. Growth rates will likely rise by at least an order of magnitude, and probably further, until we run into severe resource constraints. Just as the last 200 years have experienced more change than 10,000 BCE to 0 BCE, we are likely to see periods of 4 years in the future that experience more change than the last 200.
2. Human wages will fall, probably very far. When humans work, they will probably be improving other humans' lives (for example, in domains where we intrinsically value service by humans) rather than by contributing to overall economic productivity. The great majority of humans will probably not work. Hopefully humans will remain relatively rich in absolute terms.
3. Human values won't be the only thing shaping the future. Today humans trying to influence the future are the only goal-oriented process shaping the trajectory of society. Automating decision-making provides the most serious opportunity yet for that to change. It may be the case that machines make decisions in service of human interests, that machines share human values, or that machines have other worthwhile values. But it may also be that machines use their influence to push society in directions we find uninteresting or less valuable.

My guess is that the first two impacts are relatively likely, that there is unlikely to be a strong enough regulatory response to prevent them, and that their net effects on human welfare will be significant and positive. The third impact is more speculative, probably negative, more likely to be prevented by coordination (whether political regulation, coordination by researchers, or something else), and also I think more important on a long-run humanitarian perspective.

None of these changes are likely to occur in discrete jumps. Growth has been accelerating for a long time. Human wages have stayed high for most of history, but I expect them to begin to fall (probably unequally) long before everyone becomes unemployable. Today we can already see the potential for firms to control resources and make goal-oriented decisions in a way that no individual human would, and I expect this potential to increase continuously with increasing automation.

Most of this discussion not particularly new. The first two ideas feature prominently in Robin Hanson's speculation about an economy of human emulations (alongside many other claims); many of the points below I picked up from Carl Shulman; most of them are much older. I'm writing this post here because I want to collect these thoughts in one place, and I want to facilitate discussions that separate these impacts from each other and analyze them in a more meaningful way.

Growth will accelerate

There are a number of reasons to suspect that automation will eventually lead to much faster growth. By "much faster growth" I mean growth, and especially intellectual progress, which is at least an order of magnitude faster than in the world of today.

I think that avoiding fast growth would involve solving an unprecedented coordination problem, and would involve large welfare losses for living people. I think this is very unlikely (compare to environmental issues today, which seem to have a lower bar for coordination, smaller welfare costs to avert, and clearer harms).

Automating tech progress leads to fast growth

The stereotyped story goes: "If algorithms + hardware to accomplish X get 50% cheaper with each year of human effort, then they'll also (eventually) get 50% cheaper with each year of AI effort. But then it will only take 6 months to get another 50% cheaper, 3 months to get another 50% cheaper, and by the end of the year the rate of progress will be infinite."

In reality things are very unlikely to be so simple, but the basic conclusion seems quite plausible. It also lines up with the predictions of naive economic models, on which constant returns to scale (with fixed tech) + endogenously driven technology—> infinite returns in finite time.

Of course the story breaks down as you run into diminishing returns to intellectual effort, and once "cheap" and "fast" diverge. But based on what we know now it looks like this breakdown should only occur very far past human level (this could be the subject for a post of its own, but it looks like a pretty solid prediction). So my money would be on a period of fast progress which ends only once society looks unrecognizably different.

One complaint with this picture is that technology already facilitates more tech progress, so we should be seeing this process underway already. But we do see accelerating growth (see the section on the historical record, below), except for a historically brief period of 50–75 years. So this seems like a weak objection.

Substituting capital for labor leads to fast growth

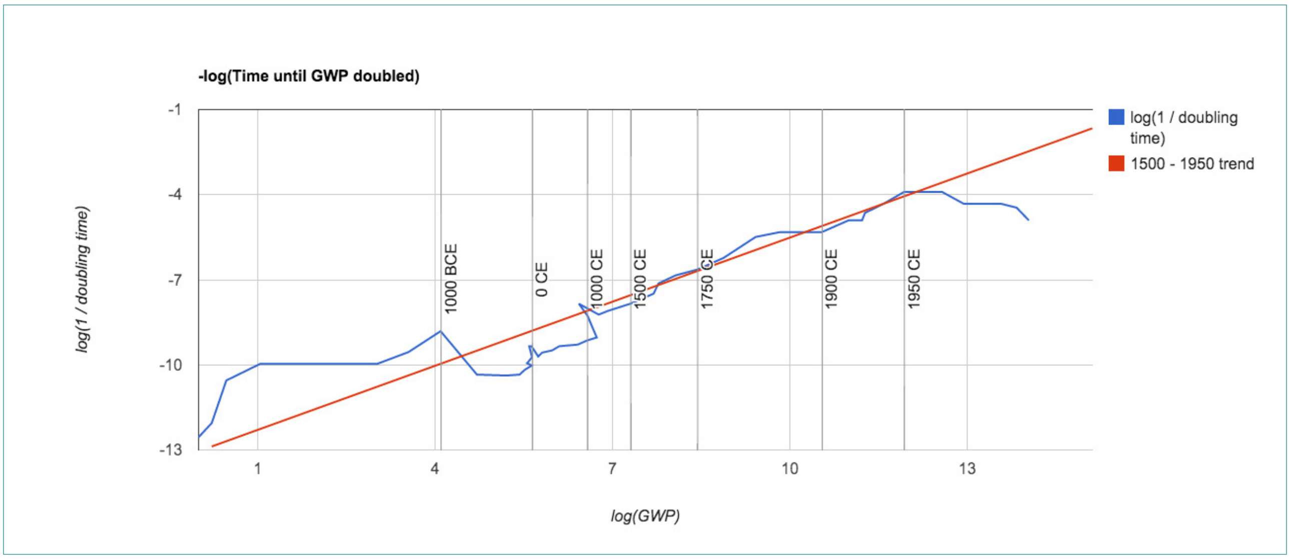
Even if we hold fixed the level of technology, automating human labor would lead to a decoupling of economic growth from human reproduction. Society could instead grow at the rate at which robots can be used to produce more robots, which seems to be much higher than the rate at which the human population grows, until we run into resource constraints (which would be substantially reduced by a reduced dependence on the biosphere).

Extrapolating the historical record suggests fast growth

Over the course of history the proportional rate of growth has increased substantially, from more than 1,500 years per doubling around 10k years ago, to around 150 years per doubling in the 17th century, to around 15 years per doubling in the 20th century. A reasonable extrapolation of the pre-1950 data appears to suggest an asymptote with infinite population sometime in the 21st century. The last 50 years represent a notable departure from this trend, although history has seen similar periods of relative stagnation.

(See the graph below, produced from Bradford Delong's data here; data and full size here. The graph starts at a doubling time of 8000 years, and plateaus around 25 years, about 300 times faster. The average time required for growth rates to double is about 1.5 doublings, or 40 years at the current rate. Each subsequent doubling of the growth rate takes half as long.)

I don't think that extrapolating this trend forward results in particularly robust or even meaningful predictions, but I do think it says one thing: we shouldn't be surprised by a future with much faster growth. The knee jerk response of "that seems weird" is inconsistent with the actual history (though it is a very reasonable intuitive reaction for someone who lived through the 2nd half of the 20th century). The more recent trend of slow growth may well continue, but I don't think we should be surprised if this stagnation was a temporary departure from trend, comparable to previous departures in severity and scale.



Wages will fall

I think this is the most robust of the three predictions. It seems very likely that eventually machines will be able to do all of the things that a human can do, as well as a human can do it. At that point, it would require significant coordination to prevent the broad adoption of machines as human replacements. Moreover, continuing to use human labor in this scenario seems socially undesirable; if we don't have to work it seems crazy to make work for ourselves, and failing to make use of machine labor would involve even more significant sacrifices in welfare.

(Humans may still demand other humans' labor in distinctively human roles. And there may be other reasons for money to move around between humans. But overall most new valuable stuff and valuable ideas in the world will be produced by machines. In the optimistic scenario, the main reason that this value will be flowing to humans, rather than merely amongst humans, will be because they own some of the machines.)

Historically humans have not been displaced by automation. I think this provides some evidence that in the near term automation will not displace humans, but in the long run it looks inevitable, since eventually machines really will be better at everything. Simple theories do suggest a regime where humans and automation are complementary, followed by a regime in which they are substitutes (as horses were once complementary with carriages, but were eventually completely replaced by automation). So at some point I think we should expect a more substantive transition. The most likely path today seems to be a fall in wages for many classes of workers while driving up wages for those who are still complementary with automation (a group that will shrink until it is eventually empty). "Humans need not apply" has been making the rounds recently, and despite many quibbles I think it does a good job of making the point.

I don't have too much to say on this point. I should emphasize that this isn't a prediction about what will happen soon, just about what will have happened by the time that AI can actually do everything humans can do. I'm not aware of many serious objections to this (less interesting) claim.

Human values won't be the only things shaping the future

I'd like to explicitly flag this section as more speculative. I think AI opens up a new possibility, and that this is of particular interest for those concerned with the long-term trajectory of society. But unlike the other sections, this one rests on pretty speculative abstractions.

Many processes influence what the world will look like in 50 years. Most of those processes are not goal-directed, and push things in a somewhat random direction; an asteroid might hit earth and kill us all, the tectonic plates will shift, we'll have more bottles in landfills because we'll keep throwing bottles away, we'll send some more radiation into space. One force stands out amongst all of these by systematically pushing in a particular direction: humans have desires for what the future looks like (amongst other desires), and they (sometimes) take actions to achieve desired outcomes. People want themselves and their children to be prosperous, and so they do whatever they think will achieve that. If people want a particular building to keep standing, they do whatever they think will keep it standing. As human capacities increase, these goal-oriented actions have tended to become more important compared to other processes, and I expect this trend to continue.

There are very few goal-directed forces shaping the future aside from human preferences. I think this is a great cause for optimism: if humans survive for the long run, I expect the world to look basically how we want it to look. As a human, I'm happy about that. I don't mean "human preferences" in a narrow way; I think humans have a preference for a rich and diverse future, that we care about other life that exists or could have existed, and so on. But it's easy to imagine processes pushing the world in directions that we don't like, like the self-replicator that just wants to fill the universe with copies of itself.

We can see the feeble beginnings of competing forces in various human organizations. We can imagine an organization which pursues its goals for the future even if there are no humans who share those goals. We can imagine such an organization eventually shaping what other organizations and people exist, campaigning to change the law, developing new technologies, etc., to make a world better suited to achieving its goals. At the moment this process is relatively well contained. It would be surprising (though not unthinkable) to find ourselves in a future where 30% of resources were controlled by PepsiCo, fulfilling a corporate mission which was completely uninteresting to humans. Instead PepsiCo remains at the mercy of human interests, by design and by necessity of its structure. After all, PepsiCo is just a bunch of humans working together, legally bound to behave in the interest of some other humans.

As automation improves the situation may change. Enterprises might be autonomously managed, pursuing values which are instrumentally useful to society but which we find intrinsically worthless (e.g. PepsiCo can create value for society even if its goal is merely maximizing its own profits). Perhaps society would ensure that PepsiCo's values are to maximize profit only insofar as such profit-maximization is in the interest of humans. But that sounds complicated, and I wouldn't make a confident prediction one way or the other. (I'm talking about firms because its an example we can already see around us, but I don't mean to be down on firms or to suggest that automation will be most important in the context of firms.)

At the same time, it becomes increasingly difficult for humans to directly control what happens in a world where nearly all productive work, including management, investment, and the design of new machines, is being done by machines. We can imagine a scenario in which humans continue to make all goal-oriented decisions about the management of PepsiCo but are assisted by an increasingly elaborate network of prosthetics and assistants. But I think human management becomes increasingly implausible as the size of the world grows (imagine a minority of 7 billion humans trying to manage the equivalent of 7 trillion knowledge workers; then imagine 70 trillion), and as machines' abilities to plan and decide outstrip humans' by a widening margin. In this world, the AI's that are left to do their own thing outnumber and outperform those which remain under close management of humans.

Moreover, I think most people don't much care about whether resources are held by agents who share their long-term values, or machines with relatively alien values, and won't do very much to prevent the emergence of autonomous interests with alien values. On top of that, I think that machine intelligences can make a plausible case that they deserve equal moral standing, that machines will be able to argue persuasively along these lines, and that an increasingly cosmopolitan society will be hesitant about taking drastic anti-machine measures (whether to prevent machines from having "anti-social" values, or reclaiming resources or denying rights to machines with such values).

Again, it would be possible to imagine a regulatory response to avoid this outcome. In this case the welfare losses of regulation would be much smaller than in either of the last two, and on a certain moral view the costs of no regulation might be much larger. So in addition to resting on more speculative abstractions, I think that this consequence is the one most likely to be subverted by coordination. It's also the one that I feel most strongly about. I look forward to a world where no humans have to work, and I'm excited to see a radical speedup in technological progress. But I would be sad to see a future where our descendants were maximizing some uninteresting values we happened to give them because they were easily specified and instrumentally useful at the time.

Potential Risks from Advanced AI

By Daniel Dewey, 2017

In this 2017 talk, Daniel Dewey presents Open Philanthropy Project's work and thinking on advanced artifical intelligence. He also gives an overview over the field, distinguishing between strategic risks – related to how influential actors will react to the rise of advanced AI systems – and misalignment risks – related to whether AI systems will reliably do what we want them to do. Watch the video by clicking on the image below.

The Open Philanthropy Project's work on Potential Risks from Advanced AI - Da...
1.00
Effective Altruism Global
2017

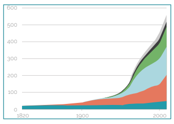
I'm the program officer at the Open Philanthropy Project in charge of potential risks from advanced AI. This is an area we're spending a lot of our senior staff time on recently, so I wanted to give an update on the work that we're doing in this area, how we think about it, and what our plans are going forward.

So, there are four basic concepts that I want to really make sure to drive home during the course of this talk, and if you watch out for these, I think they'll help you understand how we're thinking about this area.

I think there are a lot of different ways to frame potential risks from advanced AI that can inform different kinds of approaches and interventions and activities. And it can be a bit hard to understand why we're doing the things we're doing without understanding the way we're thinking about them. Also, I should mention, I didn't really frame this talk up as the perfect introduction to this area if you're not already somewhat familiar.

These are the four basic concepts:

1. Transformative AI, which is how we think broadly about the impacts that AI could have in the



Transformative AI

Strategic risks

Misalignment risks

Our strategy: Field-building

future that we care most about affecting our activities;

2. Strategic risks, having to do with how the most influential actors in the world will react to the prospect of transformative AI;
3. Misalignment risks, which have to do with being able to build AI systems that reliably do what their operators want them to do;
4. Our strategy in this area. The way we're currently planning on making a difference, which is field building.

Transformative AI

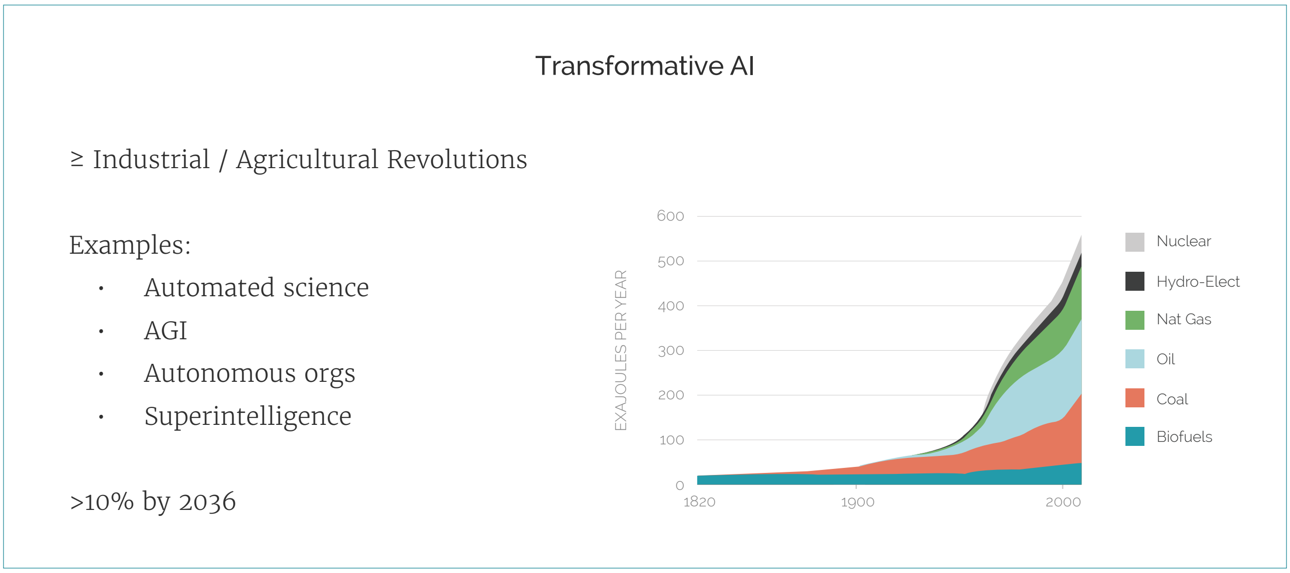
So, to start off, there's this idea of transformative AI. Basically looking ahead at the kinds of impacts we expect AI to have. We think there are a lot things that could happen and there's a lot of uncertainty about precisely what is going to happen. But something that seems reasonable is to expect AI to have an impact that is comparable to or larger than that of the Industrial or Agricultural Revolutions. And that's intended to capture a lot of possible sorts of scenarios that could happen.

So, we might see AI progress lead to automated science and technology development, which could lead to a really rapid increase in technological progress. We might see artificial general intelligence (sometimes abbreviated AGI), meaning AI systems that can do anything that a human can do, roughly. And that would really change the dynamics of the economy and how the economy functions. We might see systems that can do anything that a human or a group of humans can do. So AI systems could operate organizations autonomously. Maybe companies, non-profits, parts of government.

And then sort of looming over all of this is the idea that we shouldn't really expect AI to stop at the point of human-level competence, but we should expect the development of super-intelligent AI systems. It's not clear exactly what the distribution of capabilities of these systems would be and there are a lot of different possibilities.

The reason I've chose this picture in the slides is because it shows the change in the way human influence was wielded on the world during the Industrial Revolution. You can see this traditional set of biofuel usage down at the bottom and then over the course of the Industrial Revolution, that became a very small percentage of the overall influence that humanity wielded. Most of what we were doing in the world came to depend on these new energy sources.

The idea of transformative impact comes from AI becoming a really large percentage of how humanity influences the world. That most of the influence we have could be via AI systems that are hopefully acting



Transformative AI

≥ Industrial / Agricultural Revolutions

Examples:

- Automated science
- AGI
- Autonomous orgs
- Superintelligence

>10% by 2036

on our behalf.

Based on the conversations we've had with a lot of AI researchers, it's pretty reasonable to think that this could happen sometime in the next 20 years. I'm saying greater than 10% chance by 2036 because we said 20 years last year and so we don't want to always be saying 20 years later as years continue.

So there's this really big change in the world, there's a lot of variation in what could happen, and it's hard to predict exactly what is going to be most critical and what kinds of things we might want to make a difference on.

So here is our general strategy in this area. We can imagine two different worlds. One of them is a world where transformative AI comes somewhat by surprise, maybe it comes relatively early. And there aren't a lot of people who have been spending much of their career thinking full time about these problems, really caring about longterm outcomes for humanity. And then there's an alternate world where those professional people have existed for a while. They're working in fields with each other. They're critiquing each other's work.

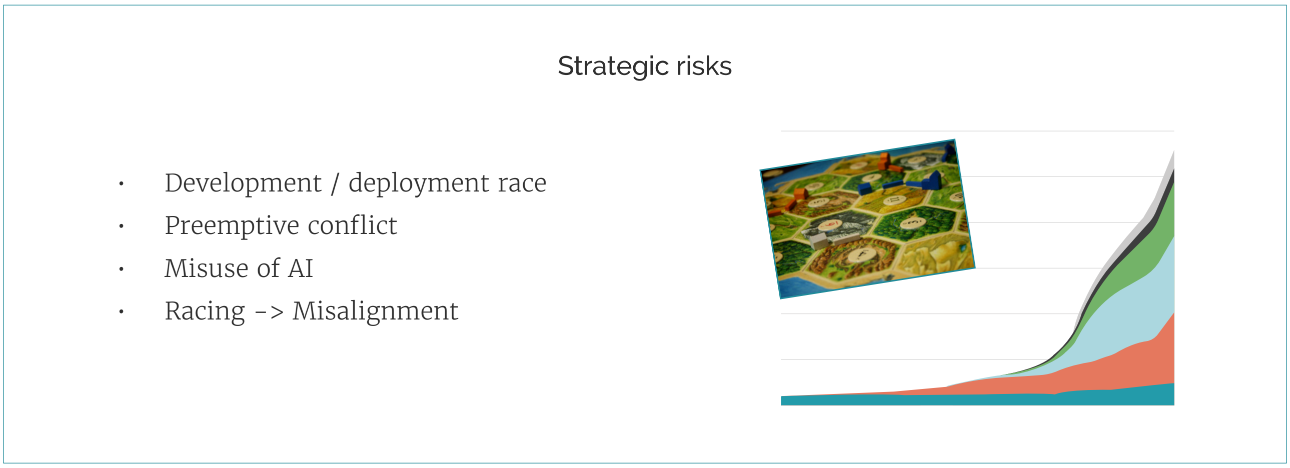
And we think that the prospect of good outcomes is a lot more likely in cases where these fields have existed for a while, where they're really vibrant. They have some of the best people in policies, some of the best people in machine learning and AI research in them. And where those people have been thinking really specifically about how transformative AI could affect the long run trajectory of human civilization.

So, our basic plan is to affect field building. To try to move these fields ahead, in terms of quality and in terms of size. And a really useful thing about this is that if you wanna affect the longterm trajectory of civilization, you don't really get to run several experiments to see which interventions are going to work well. So it's really hard to get feedback on whether what you're doing is helping.

So, what we'd like to do is start really keeping track of how these fields grow over time so that we can tell which kinds of interventions are making a difference. And it's not a sure thing that field growth is the correct strategy to pursue but it at least gives us something to measure and track to see if what we're doing is making a difference.

Strategic Risks

I'm starting with strategic risks because I think they have historically been less emphasized in the EA community. By strategic risks, I mean risks that could be caused by the way major, influential actors in the world react to the prospect of artificial general intelligence, or super-intelligence, or other kinds of transformative AI. And the way that they choose to use these technologies to affect the world. So sort of the policies and strategies they adopt.



Strategic risks

- Development / deployment race
- Preemptive conflict
- Misuse of AI
- Racing -> Misalignment

For example, if you expect this big curve of human influence in the world to be mostly about artificial intelligence in the future, then that's a big opportunity for different actors to have more influence in the future than they do today or an opportunity for that influence to be rebalanced. Maybe between different countries, between different industries. It feels like there's a strong chance that as influential actors start noticing that this might happen, that there could be preemptive conflict. There could be arms races or development races between governments or between companies.

If a government or company gains a really strong advantage in artificial intelligence, they might use it in a way that isn't in the best interest of the most people. So we could see a shift in the way resources and rights are distributed in the future. I classify that as a misuse of artificial intelligence. We want to make sure that transformative AI is used in a way that benefits the most people the most.

And then a final thing to think about is the possibility of accidental risks, risks of building AI systems that malfunction and do things that don't really benefit anyone, that weren't intentional. Then racing to develop artificial intelligence could be a big increase in that risk, because if you spend time and money and resources on making systems safer, you're spending less on racing.

What we'd like to do is build up a field of people who are trying to answer the key question of what should influential actors do in different scenarios depending on how AI development plays out. Its important to consider different scenarios because there's a lot of variation in how the future could go.

And there are a lot of existing relevant areas of expertise, knowledge and skill that seem like they're really relevant to this problem. So, geopolitics, global governance. It seems important for AI strategists to have pretty good working knowledge of AI and machine learning techniques and to be able to understand the forecasts that AI developers are making. And there's a lot of history in technology policy and the history of transformative technologies such that I hope that there are lessons that we could take from those. And of course, there's existing AI risk thought. So, Nick Bostrom's Superintelligence, things that have been done by other groups in the effective altruist community.

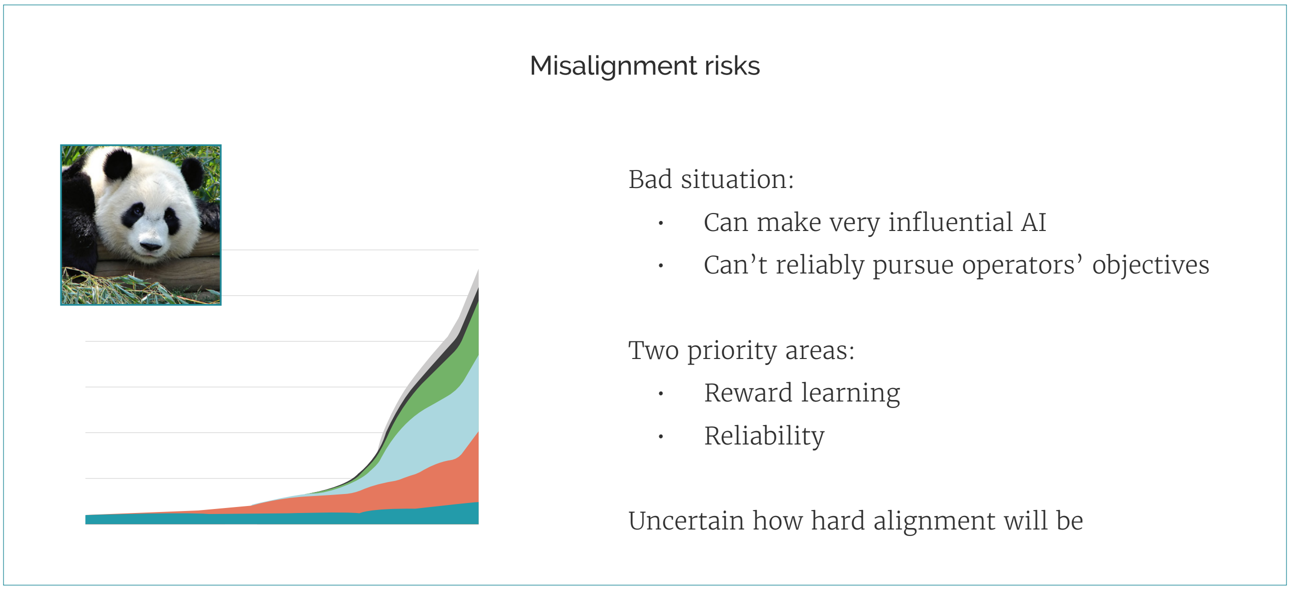
And so, our activities in this area of AI strategic risk right now, how are they going? I think that the frank summary is that we're not really sure how to build this field. Open Philanthropy Project isn't really sure. It's not really clear where we're going to find people who have the relevant skills. There's not, as far as we can tell, a natural academic field or home that already has the people who know all of these things and look at the world in this way. And so, our activities right now are pretty scattered and experimental. We're funding the Future of Humanity Institute and I think that makes sense to do, but we're also interacting a lot with government groups, think tanks, companies, people who work in technology policy, and making a few experimental grants to people in academia and elsewhere just to see who is going to be productive at doing this work.

I think it's really unclear and something I'd love to talk to people about more. Like how are we going to build this AI strategy field so that we can have professional AI strategists who can do the important work when it's most timely?

Misalignment Risks

So, the other category of risk that I wanna talk about is misalignment risks. I've used a picture of a panda. This is an adversarial example. It's a crafted image that's designed to make an AI system make an incorrect decision. And it's been sort of a recent, really hot topic in machine learning because it shows the fragility of some kinds of machine learning models that are really popular right now.

This kind of fragility is not a full picture of the problems of AI misalignment. It's not a full picture of when AI systems don't reliably do the things that their operators want them to do, but I think it's a good simple, straightforward example. The intent of training a neural network on these images was to get the neural network to make the same classifications that humans would. And it turns out to not be very hard



Misalignment risks

Bad situation:

- Can make very influential AI
- Can't reliably pursue operators' objectives

Two priority areas:

- Reward learning
- Reliability

Uncertain how hard alignment will be

to come up with a situation where the neural network will just do something completely different from what any human would say.

So, broadly speaking, misalignment risks refer to situations where we can make really influential AI systems and most of our influence over the world is flowing through these AI systems, but we can't make these systems reliably pursue the objectives that their operators intend. So, if we see this, a similar shaped graph as ended the Industrial Revolution where almost everything that humans are doing in the world is going through AI systems, and most of the way the world goes in the future depends on those decisions sort of lining up well with what humans want, then it's a really bad situation if we're not really sure if AI systems are going to do the things we want them to do, if they misinterpret what we want them to do, if they're gonna act unreliably when they're in situations we haven't anticipated before.

So, we've been talking a lot to groups like the Machine Intelligence Research Institute, to the Future of Humanity Institute, and also to technical advisors of ours who are at industrial research labs like OpenAI and Deep Mind and then also to people in academia, machine learning researchers.

And there are a couple of priority areas of research that we think are really important if you want to advance the technical capability of building AI systems that reliably do the things that their operators want them to do: reward learning and reliability.

So reward learning is this idea that it would be quite bad if we could build AI systems that can pursue easily specifiable goals like things you can measure in the world that are like how much money is in this bank account or how rewards come in through this particular channel that's flowing back to the AI. Most of the things humans care about in the world aren't easily measured in that way. So, there's a question of whether we can get AI systems to learn a task by interacting with humans in a way that makes them sort of cooperatively refine their understanding of what our goals are and act conservatively in cases where they have a lot of uncertainty and where the impact on the world could be very great if they've made the wrong evaluation of what their operator's objectives are.

And then on the reliability side, there's this question of how we train AI systems in really limited subsets of the situations that they'll eventually be functioning in. So if we want AI systems to make important decisions in the world, especially if the world is changing rapidly and dramatically, we need to be really sure that AI systems are not going to function dramatically differently in those situations than they did in training.

At Open Philanthropy Project, we've encountered a bunch of different models and ideas about how hard AI alignment will be. There's some people we've talked to who think that AI alignment is like really, really closely related to all of the things that we'll need to do in order to make AI systems effective in the world in the first place. Those problems are just gonna be solved along the way. On this view, maybe it doesn't hurt to get started ahead of time, but it's not an urgent issue. And we've talked to other people who think that there are a ton of open, unsolved problems that we have no idea how to make traction on. And that we need to get started yesterday on solving these problems. And there are a lot of people in the middle. Probably the majority of people are somewhere in between, in terms of AI and machine learning researchers.

So, we're highly uncertain about how hard alignment will be and we think that it makes a lot of sense to get started on this academic field building in this area. If the worst case scenario is that we build this field and the problems turn out to be easier than we expected, that seems pretty good.

I think we're a lot clearer how misalignment field building will go than we are about how strategic risk field building will go. In reward learning and reliability, and then in AI alignment more broadly, I think that the academic field of AI and machine learning research contains the people who have the kinds of skills and capabilities that we need for AI alignment research already. And this is an area where philanthropic funding can just directly have an impact. There's a bit of a funding puzzle to do with having all these different chickens and eggs that you need in order to get a good research field up and running. And that includes having professors who can host students, having students who are interested in working on these problems and having workshops and venues that can coordinate the research community and kind of weave people together so that they can communicate about what questions are most important.

I think it's obvious that this kind of field building work could pay off in the longer term. If you imagine this AI alignment community building up over many decades, it's obvious. But actually, I think that even if we want to develop experts who will be ready to make essential contributions on short timelines, this is among the best ways to do that, because we're finding PhD students who have a lot of the necessary skills already and getting them to start thinking about and working on these problems as soon as we can.

So, this is a scenario where we've done a pretty significant amount of grant making so far and we have some more in the works. There have been a couple big grants to senior academics in artificial intelligence and machine learning. The biggest ones being to Stuart Russell and his co-investigators, several other professors, at the Center for Human Compatible AI, which is based in Berkeley and also has branches at couple of their universities. There's another big grant that went to Joshua Bengio and bunch of his co-investigators at The Montreal Institute for Learning Algorithms. And that's a fairly recent grant. There are more students coming into that institute in the fall who we're hoping to get involved with this research.

With other professors, we're making some planning grants so that we can spend time interacting with those professors and talking with them a lot about their research interests and how they intersect with our interests in this area. Overall, we're taking a really personal, hands-on approach with grants to academic researchers in this area because I think our interests and the research problems we think are most important are a little bit unusual and a little bit difficult to communicate about.

So, I think it's important for us to do these sort of relationship-based grants and to really spend the time talking to the students and professors in order to figure out what kinds of project would be most effective for them to do.

So far, the main support that we've lent to students is via their professors. So often academic grants will support a professor, part of a professor's time and much of several of their students' times. But this fall we're hoping to offer a fellowship for PhD students, which is a major way that machine learning PhD students are supported.

I'm quite bullish on this. I think that it's reasonable to expect a lot of the really good research and ideas to come from these PhD students who will have started thinking about these things earlier in their careers and had more opportunity to explore a really wide variety of different problems and approaches. But again, offering a PhD fellowship is not something we've done before so I think it's going to be sort of experimental and iterative to figure out how exactly it's going to work.

As far as workshops, we've held a workshop at Open Philanthropy Project for a bunch of grantees and potential grantees. Basically, as an experiment to see what happens when you bring together these academics and ask them to give talks about the AI alignment problem. We were quite happy with this. I think that people quickly jumped on board with these problems and are exploring a set of ideas that are closely related to the fields that they were working on before, but are approaching them from an angle that's closer to what we think might be required to handle AI alignment.

There are also workshops like Reliable Machine Learning in the Wild that have been in academic machine learning conferences, which are the major way that academics communicate with each other and publish results. Conferences dominate over journals in the field of machine learning. So we think supporting workshops at conferences is a good way to build up this community.

And it really depends on being able to communicate these problems to professors and students because they're the primary organizing force in these workshops.

There are other developments that I think you guys might be especially interested in. There's the Open Philanthropy Project partnership with OpenAI, which I think Holden talked about a little bit yesterday. We're quite excited about this. It's an unusual grant because it's not the case that we're just contributing money to a group and then letting them pursue the activities that they were going to pursue anyway. It's like a really active partnership between us and them to try to pool our talents and resources to pursue better outcomes from transformative AI.

So, I'm really excited about that. It's not clear exactly what kinds of results and updates and communications it makes sense to expect from that because it's still pretty early, but I have high hopes for it. We funded the Machine Intelligence Research Institute last year and we're still in a lot of conversations with them about their particular outlook on this problem and the work that they're doing.

How's it going?

Professors	Students	Workshops
• Russell & co., CHAI • Bengio & co., MILA • 2 planning grants	(Supported with professors)	Open Phil, March '17 RMLW
(Hoping to add a few more soon)	(Aiming for a PhD Fellowship in late 2017)	(Will likely continue)

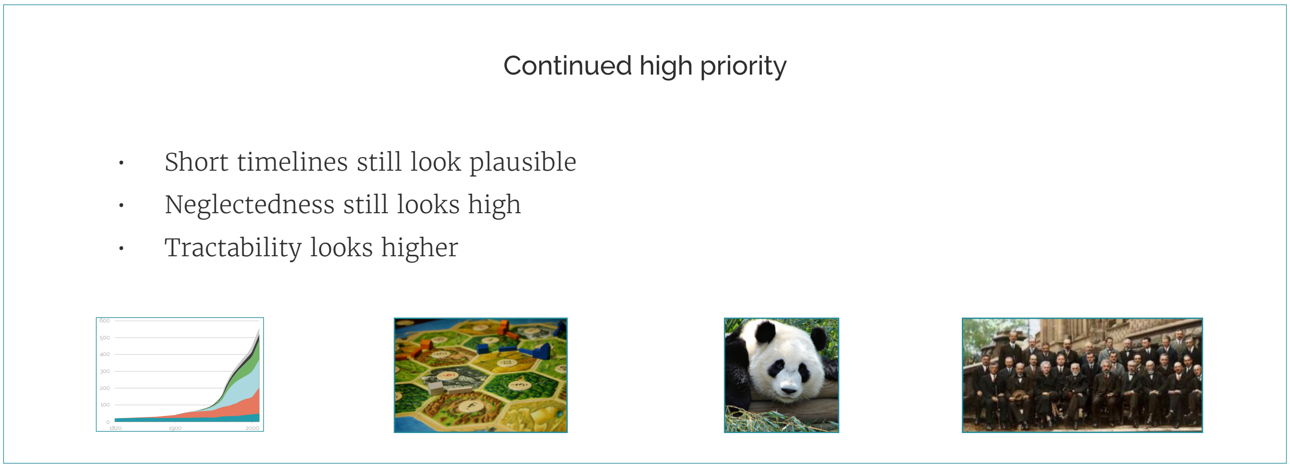
There's a collaboration between OpenAI and Deep Mind. So this is something that the Open Philanthropy Project isn't funding or playing a role in directly, but I think it's an exciting development just for people who care about this area. So, OpenAI's a nonprofit and Deep Mind is part of Google, but in theory they could be viewed as competitors for producing artificial general intelligence. So I think it's really encouraging to see their safety teams working together and producing research on the alignment problem. I think that's a robustly positive thing to do.

I also happen to think that the research that they did jointly publish, which is about learning from human feedback – so, having an AI system demonstrate a series of behaviors and having a human rate those behaviors and using those ratings to guide the learning of the AI system – I think this is a really promising research direction. A lot of this research is related to Paul Christiano's concept of act-based agents, which personally I'm really optimistic about as a new direction in the AI alignment problem.

Our Strategy in this Area

So, overall, the takeaway here: last year we published a blog post on the philanthropic opportunity that we saw from transformative AI. And looking back on that a year later, I think that short timelines still look plausible. This greater than 10% chance over the next 20 years of developing transformative AI seems really real. And additionally, we increasingly think that Open Philanthropy Project can make the biggest difference in the world where timelines are short in that way. So, a major criterion that we apply to the work that we're doing is: would this be useful if AGI were developed within the next 20 years or so.

Neglectedness still looks really high. We haven't seen a lot of other funders jumping into this space over the next year and I think it was really possible given the increase in attention to artificial general intelligence, that this space would become much more crowded. I think Open Philanthropy Project and this community are still in a pretty unusual position to influence outcomes in this area just because it is so neglected.



Continued high priority

- Short timelines still look plausible
- Neglectedness still looks high
- Tractability looks higher

And after having done some experiments in strategy and field building in technical AI alignment research, I think tractability looks higher than it did before. It's probably within the general range that we thought it was in, but maybe more concentrated in the high end. Just as we've gone on and talked to more and more AI researchers, it's been easier than expected to communicate the things that we're interested to find common ground between what they think they could do productive research on and what we think would make the biggest difference for the future trajectory of human civilization.

So those are the continued high-priorities for us. We're still spending a lot of senior staff time on it and I think it's a cause area that it makes sense to pay attention to if you're interested in the long-term trajectory of human civilization.

I'll take questions now, and thanks for your time.

Question: Do you think that we should or if it is even possible to slow the advance of AI until some of these areas can mature that you're investing in?

Daniel Dewey: I think that's a good question. My current guess is that we don't have very good levers for affecting the speed of AI development. I think there's so much money and so much pressure in the rest of society to develop artificial intelligence that it's not in a place where we have a particularly strong advantage. Slowing down technology is, I think, quite difficult to do and it would take a really concerted effort on the part of a much larger community.

But on top of that, I think it's a really open question how much it makes sense to think of this as like a race between two totally separate technologies, which are like capabilities and safety. My experience has been that you need a certain amount of capability in order to really do a lot of the research on AI safety.

So, yeah. It doesn't seem that tractable to me and even if it were more tractable, I think it's still sort of an open strategic question.

Question: Given the massive advantage that someone or some group could gain from winning the AI race, let's say, it seems to this questioner that the strategic considerations are perhaps the biggest risk. So, how does the field building that you're engaged in help us avoid this sort of arms race scenario in AI?

Daniel Dewey: I don't want to express too much confidence about this, but the way that I currently see the strategic field building work playing out is that we don't really want people making up their strategies on the fly, in a panic at the last minute. And if there are people who have done work ahead of time and gained expertise in the strategic considerations that are going on here, I think that we can have much better, more detailed, more well worked out plans for groups to coordinate with each other to achieve their shared interests.

And then also if there are some groups that we think will use AI more responsibly, or some governmental structures that we think would be more conducive to overall flourishing, I think that's not something you can work out at the last minute. So, I see developing a strategy for mitigating harms from misuse or from racing as something that we need these strategy experts to do. I don't think it's something that we can do in our spare time or something that people can do casually while they're working on something else. I think it's something that you really want people working on full time.

So I guess that's my perspective. Since we don't know what to do, that we should develop these experts.

Question: Another question that touches on several of the themes that you just mentioned there. How do you expect that AI development will impact human employment and how do think that will then impact the way that governments choose to engage with this whole area?

Daniel Dewey: Yeah. This a super good question.

I don't have a good answer to this question. I think that there are interesting lessons from self-driving cars where I think most people who have been keeping up with self-driving cars, with the raw technological progress, have been a little bit surprised by the slowness of this technology to roll out into the world.

So, I think one possibility that's worth considering is that it takes so long to bring a technology from a proof of concept in the lab to a broad scale in the world. That there could be this delay that causes a big jump in effective capabilities in the world where maybe we have, in the lab, the technology to replace a lot of human labor but it takes a long time to restructure the marketplace or to pass regulatory barriers or handle other mundane obstacles to applying a new technology.

But I think it's absolutely worth considering and it's an important strategic question if there going to be things like employment or things like autonomous weapons that will cause governments to react dramatically to AI in the really short term. In the US the big example is truck driving. Is autonomous truck driving going to cause some concerted reaction from the US government? I don't really know. I think this is a question we would like to fund to answer.

Question: Obviously, there's a lot of debate between openness and more closed approaches in AI research.

Daniel Dewey: Yeah.

Question: The grant to OpenAI's a big bet, obviously, on the open side of that ledger. How are you thinking about open and closed or that continuum between those two extremes and how does your bet on OpenAI fit into that?

Daniel Dewey: So, I don't actually think that the bet on OpenAI is a strong vote in favor of openness. I think that their philosophy, as I understand it in this area, is that openness is something that they think is a good heuristic. Like it's a good place to start from in some sense. That if one of the things you're worried about is uneven distribution of power, there's this powerful mechanism of distributing information and capabilities and technology more widely.

But if you go and look at what they've written about it, especially more recently, they've been pretty clear that they're going to be pragmatic and flexible and that if they're sitting around a table and they've developed something and their prediction is that releasing it openly would cause horrible consequence, they're not going to be like, "Well, we committed to being open. I guess we have to release this even though we know it's going to be awful for the world."

My perspective on openness is that, I mean, this is a boring answer. I think it's one of these strategic questions that like you can do a shallow analysis and say like, if you're worried about the risk of a small group of people taking a disproportionate chunk of influence and that that would be really bad, then maybe you want to be more open. If you're mostly worried about offense beating defense and only one hostile actor could cause immense harm, then you're probably gonna be more excited about closedness then openness.

But I think we need to move past this shallow strategic analysis. Like, we need people working in a real way on the detailed, nitty-gritty aspects of how different scenarios would play out, because I don't think there's a simple conceptual answer to whether openness or closedness is the right call.

Question: Well, we'll have it to leave it there for today. Round of applause for Daniel Dewey.

Daniel Dewey: Cool. Thank you.

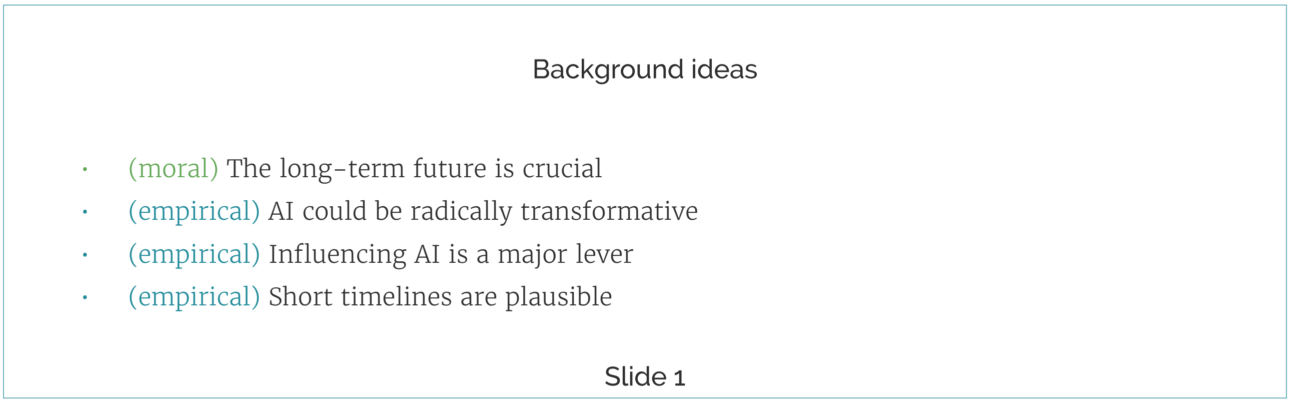
What Does (and Doesn't) AI Mean for Effective Altruism?

By Owen Cotton-Barratt, 2017

In this 2017 talk, The Future of Humanity Institute's Owen Cotton-Barratt discusses what strategy effective altruists ought to adopt with regards to the development of advanced artificial intelligence. He argues that we ought to adopt a portfolio approach – i.e., that we ought to invest resources in strategies relevant to several different AI scenarios. Watch the video by clicking on the image below.

What does (and doesn't) AI mean for effective altruism? - Owen Cotton-Barratt

Some of you may have noticed that a bunch of people in this community seem to think that AI is a big deal. I was going to talk about that a little bit. I think that there are a few different ideas which feed into what we should be paying a lot of attention to. One is that from a moral perspective, the biggest impacts of our actions – and perhaps even overwhelmingly so – are the effects of our actions today on what happens in the long term future. Then there's some pretty empirical ideas. One is that artificial intelligence might be the most radically transformative technology that has ever been developed. Then actually artificial intelligence is something that we may be able to influence the development of. Influencing that could be a major lever over the future. If we think that our actions over the long term future are important, this could be one of the important mechanisms. Then as well, that artificial intelligence and the type of radically transformative artificial intelligence could plausibly be developed in the next few decades.



Background ideas

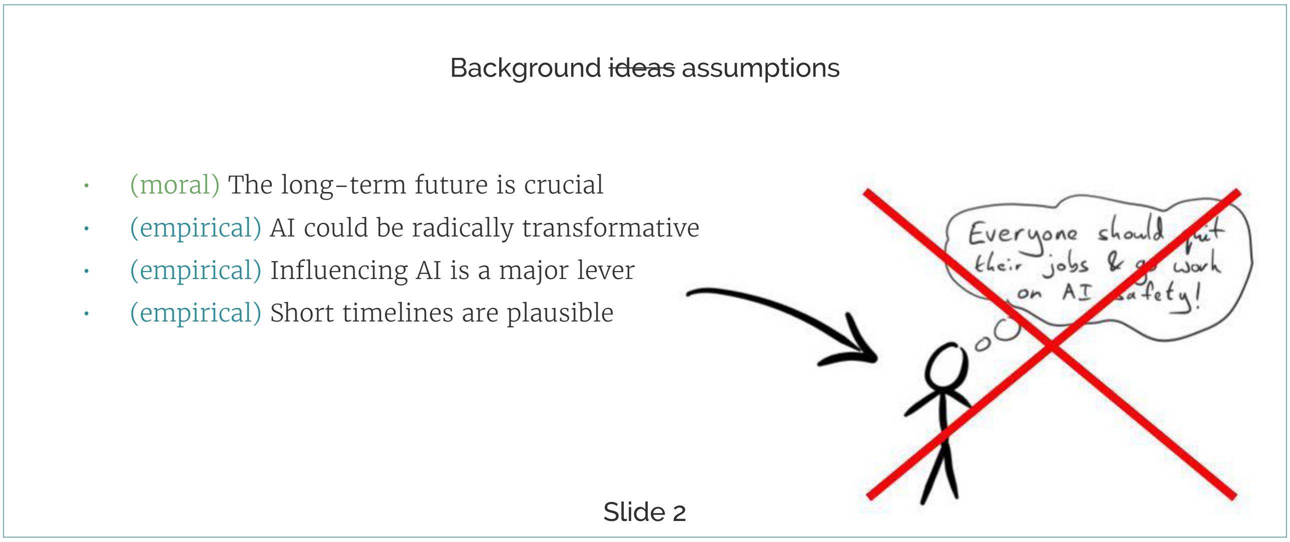
- (moral) The long-term future is crucial
- (empirical) AI could be radically transformative
- (empirical) Influencing AI is a major lever
- (empirical) Short timelines are plausible

Slide 1

I don't know what you think of all of these claims. I tend to think that they're actually pretty plausible. For the rest of this talk, I'm going to be treating these as assumptions, and I want to explore the question: if we take these seriously, where does that get us? If you already roughly agree with these, then you can just have a like sit back and see how much you agree with the analysis, and maybe that's relevant for you. If you don't agree with one of those claims, then you can treat this as an exercise in understanding how other bits of the community might think. Maybe some of the ideas will actually be usefully transferrable. Either way, if there are some of these that you haven't thought much about before, I encourage you to go and think about them – take some time afterwards. Because it seems to me at least that these are, each of these ideas is something which potentially has large implications for how we should be engaging in the world in this project of trying to help it. It seems like it's therefore the kind of thing which is worth having an opinion on.

Okay, so I'm going to be exploring where this gets us. I think a cartoon view people sometimes hold is if you believe in these ideas, then you think everybody should quit what they're working on, and drop everything, and go and work on the problem of AI safety. I think this is wrong. I think there are some related ideas in that vicinity where there's some truth. But it's a much more nuanced picture. I think for most people, it is not correct to just quit what they're doing, to work on something safety related instead. But I think it's worth understanding in what kind of circumstances it might be correct to do that, and also how the different pieces of the AI safety puzzle fit together.

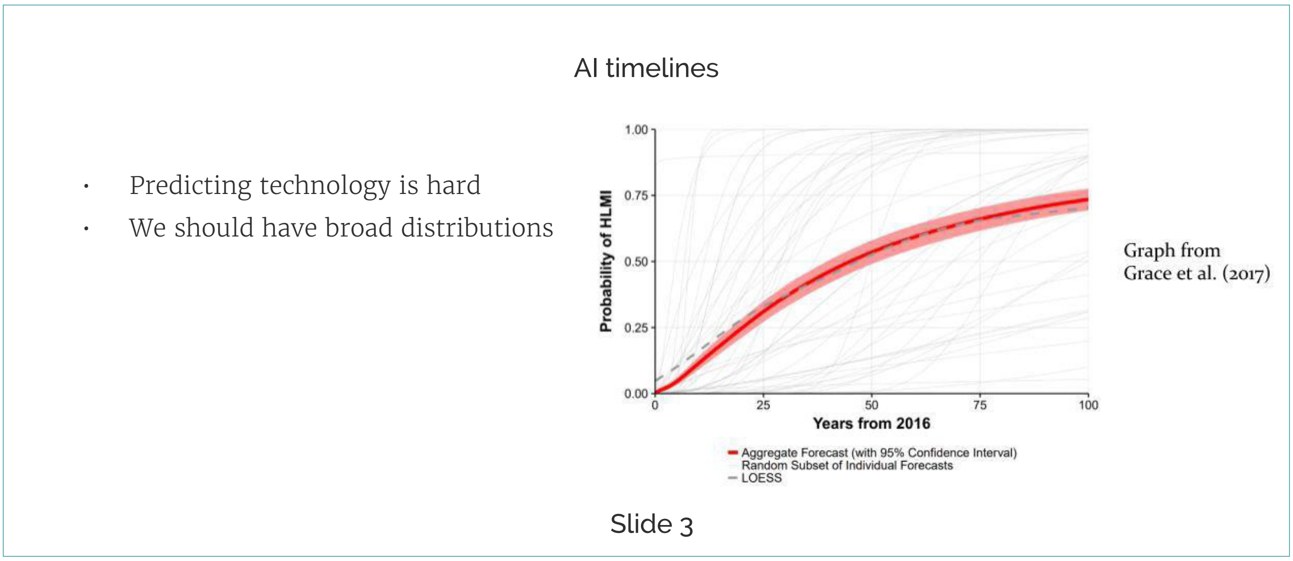
I think that thinking about timelines is important for AI. It is very hard to have any high level of confidence in when AI might have different capabilities. Predicting technology is hard, so it's appropriate



Background ~~ideas~~ assumptions

- (moral) The long-term future is crucial
- (empirical) AI could be radically transformative
- (empirical) Influencing AI is a major lever
- (empirical) Short timelines are plausible

Slide 2



AI timelines

- Predicting technology is hard
- We should have broad distributions

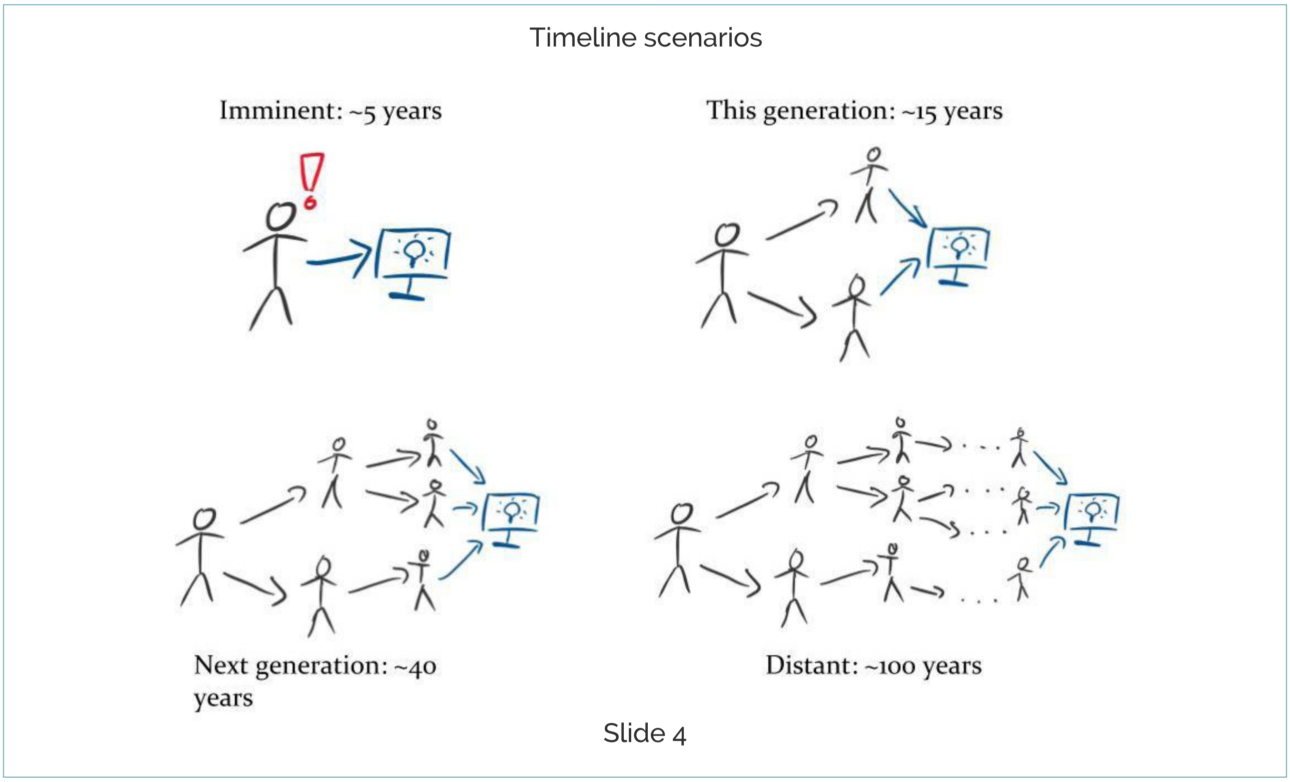
Graph from Grace et al. (2017)

Slide 3

to have uncertainty. In fact, here's a graph [Slide 3].

You can see the bunch of faint lines showing individual estimates of people working in machine learning research of when they expect high level AI to be developed. Then this bold red thing is the median of those. That's quite a lot of uncertainty. If you take almost any individual's view, and certainly this aggregate view, that represents very significant uncertainty over when transformative AI might occur. So we should be thinking about that.

Really our uncertainty should follow some kind of smooth distribution. For this talk, I'm gonna talk about four different scenarios. I think that the advantage of anchoring the possibilities as particular scenarios and treating them as discrete rather than continuous is that it becomes easier to communicate about, and it becomes easier to visualize, and think, "Okay, well what would you actually do if the timeline has this type of length?"



Timeline scenarios

Imminent: ~5 years

This generation: ~15 years

Next generation: ~40 years

Distant: ~100 years

Slide 4

The first scenario represents imminent AI, maybe something on the scale of 0 to 10 years away. In this case, it's more likely that we actually know or can make educated guesses already about who the important actors will be around the development of AI.

I want to explore a little bit about what strategies we might pursue based on each of these different timelines. If you assume this first one, then there's no time for long processes. If your idea was, "Well, I'll do a degree in CS, and then I'll go and get a PhD in machine learning, and then I'll go into research," you're too late. On the other hand, if you are already in a position where you might be able to do something in the short term, then it could be worth paying attention to. But I feel for a lot of people, even if you think there is some small chance of this first scenario happening (which in general you want to pay attention to) it may be that there isn't a meaningful way to engage.

The next possible scenario would be maybe between 10 and 25 years out. This is a timescale in which people can naturally build careers. They can go and they can learn things. They can develop networks. They can build institutions. They can also build academic fields. You can ask questions, get people motivated, and get them interested in the framing of the question that you think is important. You can also have time for some synthesis and development of relevant ideas. I think that building networks where we persuade other people who maybe aren't yet in a direct position of influence, but might be later, can be a good idea.

If we look a bit further to another possible scenario, maybe between 25 to 60 years out, that's a timescale at which people who are in the important fields today may be retiring. Paradigms in academic fields might have shifted multiple times. It becomes hard to take a zoomed in view of what it is that we need, but this means that it's more important and build things right rather than quickly. We want to build solid foundations for whatever the important fields are. When I say the important fields here, I'm thinking significantly about technical fields of how we build systems which do what we actually want them to do. I'm also thinking about the kind of governance, policy, and processes in our society around AI. Who should develop AI? How should that be structured? Who is going to end up with control over the things which are produced?

These scenrios are all cartoons. I'm presenting a couple of stylized facts about each kind of timeline. There will be a bit of overlap of these strategies, but just to give an idea of how actually the ideal strategy changes. Okay. The very distant maybe more than 60 years out, anything, maybe it's even hundreds of years, at this level predictability gets extremely low. If it takes us this long to develop radically transformative AI, it is quite likely that something else radically transformative will have happened to our society in the meanwhile. We're less likely to predict what the relevant problems will be. Instead, it makes sense to think of a strategy of building broad institutions, which are going to equip the people of that time to better deal with the challenges that they're facing then.

I think actually it's plausible that the effective altruism community, and the set of ideas around that community, might be one broad, useful institution for people of the far future. If we can empower people with tools to work out what is actually correct, and the motivation and support to act on their results, then I'd be like, "Yep, I think we can trust those future people to do that."
The very long term is the timescale at which other very transformative things occurring in our society are

more likely to happen. This can happen on the shorter timescales as well. But if you think on a very long timescale, there is much more reason to put more resources toward other big potential transitions, rather than just AI. I think that AI could be a big deal, but it's definitely not the only thing that could be a big deal.

Okay. I've just like talked us through different timelines. I think that most reasonable people I know put at least some nontrivial probability on each of these possible scenarios. I've also just outlined how we probably want to do different things for the different scenarios. Given all of this, what should we actually be doing? One approach is to say, "Well, let's not take these on the timelines. Let's just do things that we think are kind of generically good for all of the different timelines." I think that that's a bad strategy because I think it may miss the best opportunities. There may be some things which you only notice are good if you're thinking of something more concrete rather than just an abstract, "Oh, there's gonna be AI at some point in the future." Perhaps for the shorter timelines, that might involve going and talking to people who might be in a position to have any effect in the short term, and working out, "Can I help you with anything?"

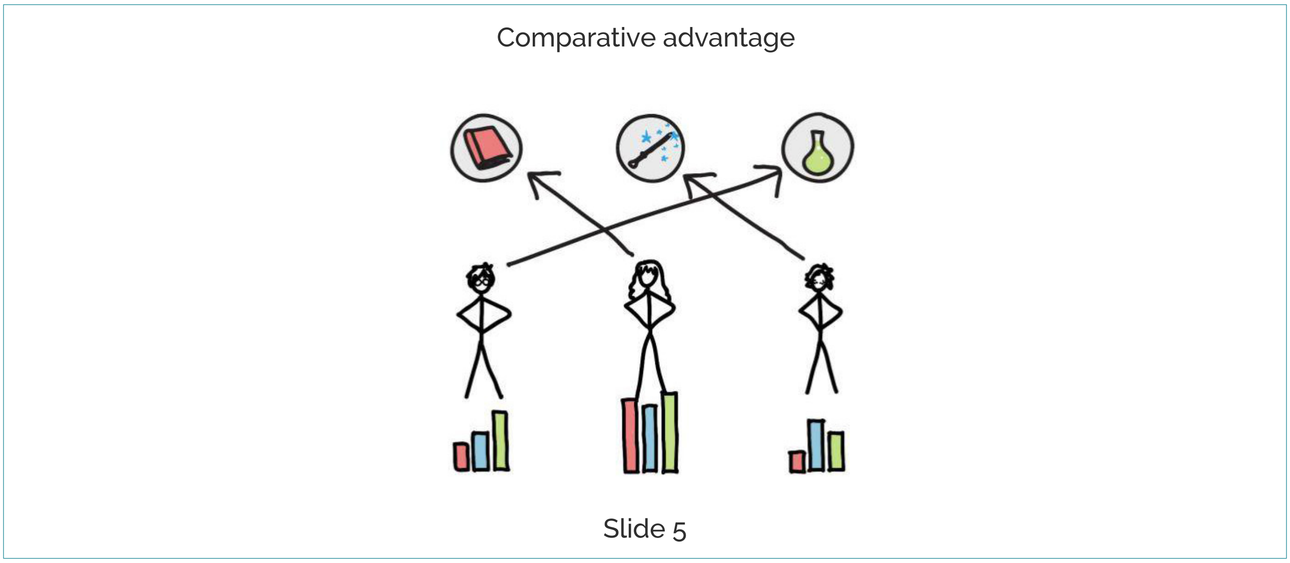
Okay. The next kind of obvious thing to consider is, well, let's work out which of these scenarios is the most likely. But if you do this, I think you're missing something very important, which is that we might have different degrees of leverage over these different scenarios. The community might have different leverage available for each scenario. It can also vary by individual. For the short timelines, probably leverage is much more heterogeneous between different people. Some people might be in a position to have influence, in that case it might be that they have the highest leverage there. By leverage, I mean roughly, conditional on that scenario actually pertaining, how much does you putting in a year of work, trying your best, have an effect on the outcome? Something like that.

Okay. Maybe we should just be going for the highest likelihood multiplied by leverage. This of course is like the place where we have the most expected impact. I think there's something to that. I think that if everybody properly does that analysis for themselves and updates as people go and take more actions in the world, then in theory that should get you to the right things. But the leverage of different opportunities varies both as people take more opportunities and also even just for an individual. I've known people who think that they've had different opportunities they can take to help short timelines and then a bunch of other opportunities to help with long timelines. This is a reason not to naively go for highest likelihood multiplied by leverage.

Okay. What else? Well, can we think about what portfolio of things we could do? I was really happy about the theme of this event because thinking about the portfolio and acting under uncertainty is something I've been researching for the past two or three years. On this approach, I think we want to collectively discuss the probabilities of different scenarios, the amount of leverage we might have for each, and the diminishing returns that we have on work aimed at each. Then also we should discuss about what that ideal portfolio should look like. I say collectively because this is all information where when we work things out for ourselves, we can help inform others about it as well, and we can probably do better using collective epistemology than we can individually.

Then we can individually consider, "Okay, how do I think in fact the community is deviating from the ideal portfolio? What can I do to correct that?" Also, "What is my comparative advantage here?" Okay. I

want to say a couple of words about comparative advantage. I think you know the basic idea. Here's the cartoon I think of it in terms of:
You've got Harry, Hermione, and Ron, they have three tasks to do, and they've gotta do one task each.



Comparative advantage

Slide 5

Hermione is best at everything, but you can't just get Hermione to do all the things. You have to allocate them one to one. So it's a question of how do you line the people up to the things so that you have everyone doing something that they're pretty good at it, and overall you get all of the important things done? I think that this is something that we can think of at the level of individuals choosing, "What am I going to work on? Well, I've got this kind of skillset." It's something that we can think of at the level of groups as well. We can ask, "What is my little local community going to work on?" or "What is this organization going to do, and how do we split up responsibility between different organizations?"

Comparative advantage is also a concept you can think of applied over time. This is a little bit different because people's actions in the past are fixed, so we can't affect those. But you can think there's things that might want to be done and we can do some of these. People in the past did some of them. People in the future might do some of them and there's a coordination question of what we have a comparative advantage at relative to people in the future. This is why when I was looking at longer scenarios, the next generation in the distant cases, I was often thinking it was better to let people in the future solve the concrete problems. They're gonna be able to see more clearly what is actually to be solved. Meanwhile, we have a comparative advantage at building the processes, the communities, the institutions which compound over time, and where getting in early is really helpful.

If you're taking something like this portfolio approach, I think that most projects should normally have at least a main scenario in mind. This forces you to be a little bit more concrete and to check that the things you're thinking of doing actually line up well with the things which are needed in some possible world. I also think you want to be a bit careful about checking that you're not doing anything which would be bad for other scenarios. There's always an opportunity cost. If you're doing something where you're thinking, "I want to help with this short timeline scenario," then you're not doing something else you could've done to help with the next generation in a longer timeline scenario.

You could also have situations where maybe I would think that if AI is imminent, the right thing to do is to run around and say, "Everybody panic. AI is coming in five years. It's definitely coming in five years."

If it definitely were coming in five years, maybe that would be the right thing to do. I actually don't think it is. Even if it were, I think that would be a terrible idea because if you did that, then people, if it didn't occur in five years and we were actually in a world where radically transformative AI was coming in 25 years, then in 15 years, a lot of people are gonna go, "We've heard that before," and not want to pay attention to it. This is a reason to have an idea of paying some attention to the whole idea of the portfolio that as a community we want to be paying attention to even if individually, most projects should have a main scenario in mind. Maybe as an individual, your whole body of work has a main scenario in mind. It's still worth having an awareness of where other people are coming from, and what they're working on, and what we're doing collectively then.

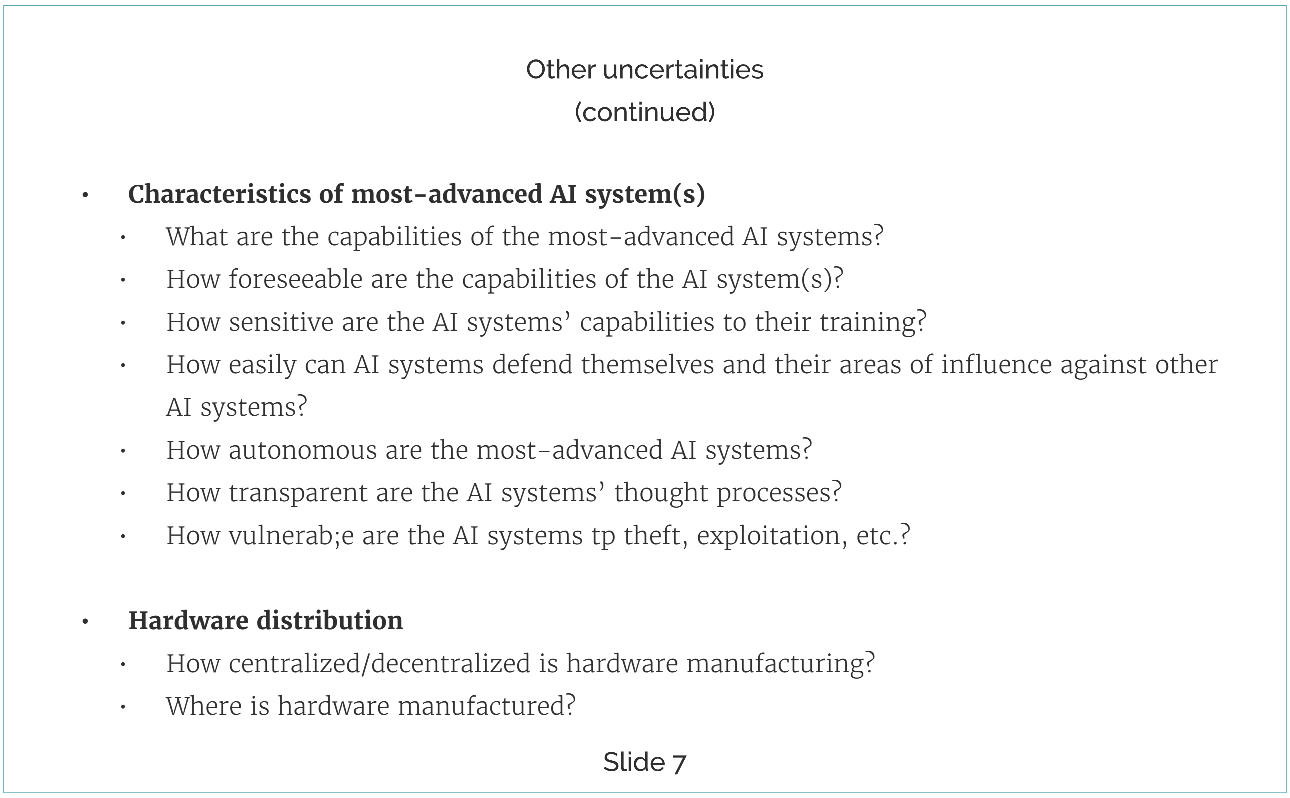
I've mostly talked about timelines here. I think that there are some other significant uncertainties about AI. For instance, how much is it that we should be focusing on trying to reduce the chances of catastrophic accidents from powerful AI? Or how much of the risk is coming from people abusing powerful technologies? We hypothesized it was gonna be a radically transformative technology with influence over the future. How much of that influence actually comes through things which are fairly tightly linked to the AI development process? Or how much influence appears after AI is developed? If most of the influence comes from what people want in the world after an AI is developed, it might makes sense to try to affect people's wants at that point.

In both of these cases, I think we might do something similar to portfolio thinking. We might say, "Well, we've put some weight on each of these possibiltiies," and then we think about our leverage again. Maybe

Other uncertainties

- **Timelines:**
    - When is high-level AI developed?
    - How quick is the transition from high-level AI to something much more powerful?
- **Development path**
    - How continuous/discontinuous is AI progress?
    - How much advance warning is there that high-level AI (or something much more powerful) is nearby?
    - Is high-level AI first reached through an AGI-like system?
    - How centralized/decentralized is deployment of high-level AI?
- **Competitive landscape**
    - How many research groups are at the forefront of AI development?
    - What types of research groups are at the forefront of AI development, e.g. industry, academic, government, coalition of governments, private-public parthership?
    - How closely competitive are the leading research groups?
    - Is there an AI "arms race"?
    - How open is AI research?
    - Is there a significant first-mover advantage?

Slide 6



Other uncertainties
(continued)

- **Characteristics of most-advanced AI system(s)**
  - What are the capabilities of the most-advanced AI systems?
  - How foreseeable are the capabilities of the AI system(s)?
  - How sensitive are the AI systems' capabilities to their training?
  - How easily can AI systems defend themselves and their areas of influence against other AI systems?
  - How autonomous are the most-advanced AI systems?
  - How transparent are the AI systems' thought processes?
  - How vulnerab;e are the AI systems tp theft, exploitation, etc.?

- **Hardware distribution**
  - How centralized/decentralized is hardware manufacturing?
  - Where is hardware manufactured?

Slide 7

for some of them, we shouldn't be split. Some of them we might do. We can't do this with all of the uncertainties. There are a lot of uncertainties about AI.

Here's a slide from another talk. It just lists a lot of questions. A lot of them about how AI might develop. We can all have nuanced views about each of these questions. That's fine. We need to do some picking and choosing here. But I do think that we should strive for nuance. I think the reason is that there's a lot of uncertainty, and we could potentially have extremely nuanced views about a lot of different things. The world is complicated, and we have a moderately limited understanding of it. One of the things which may make us better equipped for the future is trying to reduce our limits on our understanding.

What can individuals do? I think consider personal comparative advantage. You can ask yourself, "Could I seriously be a professional researcher in this?" Check with others as well. I think people vary in their levels of self-confidence, so I actually think that others' opinions often can be more grounding than our own opinion for this. It's a pretty specialized skillset that I think is useful for doing technical safety research. Most people in the community are not gonna end up with that skillset and that's fine. They should not be quitting their jobs, and going to try and work on safety research. They could be saying, "Well, I want to give money to support this," or they could be aiming at other parts of this portfolio. They could say, "Well, I want to help develop our institutions to build something where we're gonna be better placed to deal with some of the longer timeline scenarios."

You could also diversify around those original assumptions that I made. I think that each of them is pretty likely to be true. But I don't think we should assume that they are all definitely true. We can check whether in fact there are worlds where they're not true that we want to be putting some significant weight onto. I think also just helping promote good community epistemics is something that we can all play a part in. By this I mean pay attention to why we believe things and communicate our real reasons to people. Sometimes you believe a thing because of a reason like: “Well, I read this in a blog post by Carl Shulman, and he’s really smart.” He might provide some reasons in that blog post, and I might be able to pallet the reasons a little bit. But if the reason I really believe it is I read that, that’s useful to communicate to other people because then they know where the truth is grounded in the statements I’m making, and it may help them to be able to better see things for themselves, and work things out. I also think we do want to often pay attention to trying to see the underlying truth for ourself. Good community epistemics is one of these institutions which I think are helpful for the longer timelines, but I think they’re also helpful for our community over shorter periods. If we want to have a portfolio, we are going to have to coordinate and exchange views on what the important truths are.

What should individuals do?

- Consider personal comparative advantage
- Diversify re. assumptions
- Help promote good community epistemics

Slide 8

What does AI mean for effective altruism? My view is that it isn’t the one thing that everyone has to pay attention to, but it is very plausibly a big part of this uncertain world stretching out in front of us. I think that we collectively should be paying attention to that and working out what we can do, so we can help increase the likelihood of good outcomes for the long term future.

Biosecurity as an EA Cause Area

Claire Zabel, 2017

In this 2017 talk, the Open Philanthropy Project's Claire Zabel talks about their work to mitigate Global Catastrophic Biological Risks. She also discusses what effective altruists can do to help, as well as differences between biological risks and risks from advanced AI. At the very end you will find an added section on what you can do to help.

Today I'm going to talk to you about biosecurity as a cause area and how the Open Philanthropy Project is thinking about it. I think that this is a cause area that EAs have known about for a while, but haven't dug into as deeply as some of the other things that have been talked about at this event – like AI and farm animal welfare and global poverty – but I think it's important. I think it's an area where EAs have a chance to make a huge difference, especially EAs with a slightly different set of skills and interests than those required by some of the other cause areas. I would love to see more EAs engaging with biosecurity.

When I say biosecurity, I want to make sure we're clear on the problem that I'm talking about. I'm focusing on what we're calling Global Catastrophic Biological Risks at the Open Philanthropy Project. I'm going to talk to you about how we see that risk and where we see that risk – where we think it might be coming from. I'm going to talk to you about how I think EAs can make a difference in it. Then I want to note that I'm not really focusing too much on the specific work that we've done and that others have done. I thought it would be more interesting for people to get a sense of what this area is like and the strategic landscape as we see it before getting into the details of specific organizations and people, so hopefully that's helpful for everyone.

I also want to note quickly that I think this is an area where a lot less thinking has been done for a much shorter period of time, so to a greater extent everything should be viewed as somewhat preliminary and uncertain. We might be changing our minds in the near future.

The cause for concern when we think about global catastrophic biological risks is something that could threaten the long term flourishing of human civilization, that could impair our ability to have a really long, really big future full of joy and flourishing for many different sentient beings. That's kind of different from what you might think about biological risks that most people talk about, which are often things like Ebola or Zika. Ebola or Zika are unbelievably tragic for the people afflicted by them, but it doesn't seem like the evidence suggests that they have a realistic chance of causing international civilizational collapse and threatening our long-term future.

To take this further, we predict that we would need a really extremely big biological catastrophe to threaten the long-term future. We're really thinking about something that kills or severely impairs a greater proportion of the entire human civilization than what happened in either of the world wars or in the 1918 flu pandemic. That kind of implies that we're thinking about fatalities that could range into the hundreds of millions or even the billions. There's a lot of really amazing work that could go into preventing smaller risks, but that's not really what we've been focusing on so far. It's not what I anticipate us focusing on in the future. Overall, we're currently ranking the prevention of global catastrophic biological risks as a high priority, although I think it's somewhat uncertain. I think it's high priority to figure out more and then we might readjust our beliefs about how much we should prioritize it.

So what are these risks even like? We think the biggest risks are from biological agents that can be easily transmitted that can be released in one area and spread, as opposed to something like Anthrax, which is very terrible in the space that it's released, but it's hard to imagine it really coming to afflict a large proportion human civilization.

Then within the space of infectious diseases, we're thinking about whether the most risky type of attack would be something that happened naturally that just came out of an animal reservoir, or something that was deliberately done by people with the intention of causing this kind of destruction. Or it might be the middle ground of something that might have been accidentally released from a laboratory where people were doing research.

Our best guess right now is that deliberate biological attacks are the biggest risk. Accidental risk somewhere in the middle, and natural risk is low. I want to explain why that is because I think a lot of people would disagree with that. Some of the reasons I'm skeptical of natural risks are that first of all, they've never really happened before. Humans have obviously never been caused to go extinct by a natural risk, otherwise we would not be here talking. It doesn't seem like human civilization has come close to the brink of collapse because of a natural risk, especially in the recent past.

You can argue about some things like the Black Death, which certainly caused very severe effects on civilization in certain areas in the past. But this implies a fairly low base rate. We should think in any given decade, there's a relatively low chance of some disease just emerging that could have such a devastating impact. Similarly, it seems like it rarely happens with nonhuman animals that a pathogen emerges that causes them to go extinct. I know there's one confirmed case in mammals. I don't know of any others. This scarcity of cases also implies that this isn't something that happens very frequently, so in any given decade, we should probably start with a prior that there's a low probability of a catastrophically bad natural pathogen occurring.

Also, we're in a much better situation than we were in the past and than animals are in some ways, because we have advanced biomedical capabilities. We can use these to create vaccines and therapeutics and address a lot of risks from different pathogens that we could face.

Then finally, on kind of a different vein, people have argued that there's some selection pressure against a naturally emerging highly virulent pathogen because when pathogens are highly virulent, often their hosts die quickly and they try to rest before they die and they're not out in society spreading it the way you might spread the cold, if you go to work when you have the cold.

Now, before you become totally convinced about that, I think that there's some good countervailing considerations to consider about humanity, that make it more likely that a natural risk could occur now than in the past. For example, humanity is much more globalized, so it might be the case that in the past there were things that were potentially deadly for human civilization, but humans were so isolated it didn't really spread and it wasn't a huge deal. Now everything could spread pretty much around the globe.

Also, civilization might be more fragile than it used to be. It's hard to know, but it might be the case that we're very interdependent. We really depend on different parts of the world to produce different goods and perhaps a local collapse could have implications for the rest of the globe that we don't yet understand.

Then there's another argument one can usually bring up, which is if you're so worried about accidental or engineered deliberate attacks, there's also not very much evidence of those being a big deal. I would agree with this argument. There haven't been very many deliberate biological weapon attacks in recent times. There's not a strong precedent. Nonetheless, our best guess right now is that natural risks are pretty unlikely to derail human civilization.

When we think in more detail about where catastrophic biological attack risks come from, we can consider the different potential actors. I don't think that we've really come to a really strong view on this. I do want to explain the different potential sources. Some possible sources could be different states. For example, in bio-weapons programs, states could develop pathogens as weapons that have the potential to be destructive. Small groups, such as terrorists or other extremists might be interested in developing these sorts of capabilities. Individuals who have an interest, people working in various sorts of labs: in academia, in the government and on their own. There are DIY biohacker communities that do different sorts of biological experimentation. Those are the different groups that might contribute to catastrophic biological risk.

There are different kinds of pathogens and I think here – our thinking is even more preliminary – we're especially worried about viral pathogens, because there's proven potential for high transmissibilities and lethality among viruses. They can move really fast. They can spread really fast. We have fewer effective countermeasures against them. We don't have very good broad spectrum antivirals that are efficacious and that means that if we had a novel viral pathogen, it's not the case that we have a huge portfolio of tools that we can expect to be really helpful against it.

I've created this small chart that I think can help illustrate how we divide up these risks. On the top there's a dichotomy of whether the pathogen is more natural or more engineered and then on the vertical axis a dichotomy of whether it emerged naturally (or accidentally) or was a deliberate release.

General Risk Factors: Comparing Between Cases

	Naturally-arising pathogen	Engineered strain
Accidentally released or naturally emerging	Lower risk	Somewhere in-between
Deliberate release	Somewhere in-between	Higher risk

**Don't have a strong view about relative risk from top right vs. bottom left quadrant, but expect risk level is not the same or similar*

The reason I'm flagging these quadrants is because I think there are two different ways to increase the destructiveness of an attack. One is to engineer the pathogen really highly, and the other is to optimize the actual attack type. For example, if you released a pathogen at a major airport, you would expect it to spread more quickly than if you released it in a rural village. Those are two different ways in which you can become more destructive, if you're interested in doing that. Hopefully you're not. My current guess is that there's a lot more optimization space in engineering the actual pathogen than in the release type. There seems to be a bigger range, but we're not super confident about that.

Characterizing the risk: Increasing Risk
from Engineered Pathogens

- Advances in gene-editing technology (e.g. CRISPR)
- Decreasing cost of DNA and RNA synthesis
- Biotech capabilities becoming more widely available
- Technological advancement appears to be outpacing norms and regulations

Here's where the risk we see is coming from. There are advances in gene editing technology, which is a really major source of risk. I think that they've created a lot more room to tinker with biology in general to both lower resources and lower levels of knowledge required and to a greater overall degree, create novel pathogens that are different from what exists in nature, that you can understand how they work. This has amazing potential to do a lot of good but it also has potential to be misused. It's becoming a lot cheaper to synthesize DNA and RNA, to get genetic material for different pathogens. This means that these capabilities are becoming more widely available, just because they're cheaper. Regulating them and verifying buyers is becoming a bigger proportion of the costs, which means companies are more and more incentivized to stop regulating sales and verifying buyers.

Biotech capabilities are becoming more available around the world. They're spreading to different areas, to new labs. Again, this is mostly a sign of progress. People are having access to technology, places in Asia and around the world are having large groups of very talented scientists and that's really great for the most part, but it means there are more potential sources of risk than there were in the past.

Then finally, all of those things are happening much faster than governments can possibly hope to keep up and than norms can evolve, so that leads you to the situation where the technology has outpaced our society and our means of dealing with risk, and that increases the level of danger.

Now I'll contrast and compare biosecurity with AI alignment, because I think AI alignment is something people are much more familiar with. It might be helpful to draw attention to the differences, for getting people up to speed. I think that overall, there's a smaller risk of a far future negative trajectory change from biosecurity. Overall it seems like a smaller risk to me.

With addressing biosecurity risk, there are fewer potential upsides. With an AI, you can imagine that if it develops really well, it has this amazing potential to increase human capabilities and cause human

Characterizing the risk: Strategic Differences with Work On AI Alignment

- Best guesses:
  - Smaller overall risk of negative far-future trajectory change
  - Less upside expected from our program, relative to the status quo
  - More of the risk comes from bad actors
  - Many potential sources of risk and risk reduction
    - Unilateral implementation may:
      - Have relatively low resource requirements
      - Become accessible to many actors
    - But unilateral development + deployment of countermeasures is feasible

flourishing. With biosecurity, we're basically hoping that just nothing happens. The best outcome is just nothing. No attacks occur. No one dies. Society progresses. In the case of AI alignment, maybe somebody develops an aligned AI, which would be great. But for biosecurity, we're really about preventing downside risks. More of the risk here comes from people with actively bad intentions as opposed to people with good intentions or people who are just interested in the research, especially if you believe and you agree with me that deliberate attacks are the most likely source of concern.

In biosecurity more than AI, I think there are many more relevant actors on both sides, as opposed to there being a few labs with a lot of capabilities in AI. It could be the case that we end up with a situation in biosecurity where there are millions of people that are capable of doing something that would be pretty destructive. And also, we can unilaterally develop counter measures against their attacks. There's less connection between the sources of the risk and the sources of the risk reduction. They're more divorced from one another. There's more possible actors on the sides of attack and defense.

I think that the way that The Open Philanthropy Project is seeing this field right now is somewhat different from how most people are seeing it. Most of the discussion in the field of biosecurity is focused on much smaller risks than the ones that we're worried about. I think discussion of things with greater than one million fatalities was kind of taboo up until very recently. It's been difficult for us to find people that are interested in working on that kind of thing. I think that part of the reason for that, is that it's been really hard to get funding in the space, so people want to make sure their work seems really relevant. And since small attacks and small outbreaks are more common, a good way to make your work more relevant is to focus on those.

There's ongoing debate in the field about whether natural, deliberate or accidental releases are the biggest risks. I don't think people are synced up on what the answer to that question is. I don't think everyone agrees with us that deliberate is mostly the thing to worry about. Then people are really trying to walk this tightrope of regulating risky research while not regulating productive research, maintaining national competitiveness, and encouraging productive biotech R&D.

Given all of that, we have some goals in this space. They're kind of early goals. They won't be sufficient on their own. They're mostly examples, but I think they could get us pretty far. The first thing is we just really need to understand the relevant risks in particular. I'm keeping it very high level for now, because there's not a lot of time, and partly because I think that talking about some of these risks publicly is not a productive thing to do, and also because we're pretty uncertain about them. I think it would be really helpful to have some people dig into the individual risks. Think about what one would need to do in order to pull off a really catastrophic bio attack. How far out is that being a possibility? What sorts of technological advancements would need to occur? What sorts of resources would one need to be able to access in order to do that? If we can answer these questions, we can have a sense of how big catastrophic biosecurity risks are and how many actors we need to be worried about.

Understanding particular risks will help us prioritize things we can do to develop counter measures. We want to support people in organizations that increase the field's ability to respond to global catastrophic biological risks. The reason for that is that right now the field of biosecurity has lacked funding for a long time. A lot of people have left the field. Young people are having a very difficult time going into the field. Hopefully that's changing, but it's still a pretty dire situation, in my view. We want to make sure that the field ends up high quality with lots of researchers that care about the same risks we care about, so people that show signs of maybe moving in that direction, we're very enthusiastic about supporting, in general.

Then finally, we want to develop medical counter measures for the things that we're worried about. We've started having our science advisors look into this. We have some ideas about what the worst risks are and if we can develop counter measures in advance and stockpile those, I think we would be much better prepared to address risks when they come up.

Finally, I want to talk to you a little bit about what I think EAs can do to help. I see a lot of potential value in bringing parts of the EA perspective to the field. Right now there aren't a lot of EAs in biosecurity and I think that the EA perspective is kind of special and has something special to offer people. I think some of the really great things about it are, first of all, the familiarity with the idea of astronomical waste and the value of the far future. That seems like it's somewhat hard to understand. It's a bit weird and counterintuitive and philosophical, but a lot of EAs find it compelling. A lot of other people find it wacky or haven't really heard about it. I think having more concern about that pool of value and those people in the future who can't really speak for themselves could do the field of biosecurity a lot of good.

Another thing that I think is amazing about the EA perspective, is comfort with explicit prioritization, the ability to say, "We really need to do X, Y, and Z. A, B, and C are lower priority. They'll help us less. They're less tractable. They're more crowded. We should start with these other things." I think right now, the field doesn't have a clear view about that. There's not a very well thought out and developed road map to addressing these concerns. I think EAs would be good at helping with that.

Finally, I think a lot of EAs have a skepticism with established methods and expertise. That's great because I think that's necessary actually in almost every field. Especially in fields that involve a complicated interplay of natural science and social science. I think that there's a lot of room for things to be skewed in certain directions. I haven't seen too much harmful skew, but guarding against it would be really helpful.

There's some work going on at the Future of Humanity Institute that we're very excited about. It seems like there's a lot of low hanging fruit right now. There are a lot of projects that I think an EA could take on and they'd be pretty likely to make progress. I think biosecurity progress is more of a matter of pulling information together and analyzing it, and less based only in pure insight.

I think that you should consider going into biosecurity if you are an EA concerned with the far future, who wants to make sure that we all get to enjoy our amazing cosmic endowment, and if you think that you might be a good fit for work in policy or in the biomedical sciences.

This is an area where I think that a lot of safety might come from people not overhyping certain sorts of possibilities as they emerge, at least until we develop counter measures. It's important to have people that feel comfortable and are okay with the idea of doing a lot of work and then not sharing it very widely and actually not making it totally open, because that could actually be counterproductive and increase risk. That's what I hope that people will be willing to do. I hope that we find some EAs who want to move into this field. If you feel like you're interested in moving into this field, I would encourage you to reach out to me or grab me sometime at this conference and talk about both what you'd like to do and what might be stopping you from doing it.

In the future we might write more about how we think people can get into this field and be able to do helpful research, but we haven't really done that yet, so in the meantime, I really hope that people reach out. Thank you so much and I'll take your questions.

Question: Okay, so we've got a number of questions that have come in and I'm just gonna try to rifle through them and give you a chance to answer as many as we can. You emphasized the risk of viral pathogens. What about the, I think, more well known if not well understood problem of antibiotic resistance? Is that something that you're thinking about and how big of a concern is that for you?

Claire Zabel: Yeah. I think that's a good question. The Open Philanthropy Project has a report on antibiotic resistance that I encourage you to read if you're curious about this topic. I think it's a really big concern for dealing with conventional bacterial pathogens. Our best guess is that it's not such a special concern for thinking about global catastrophic biological risks, first of all, because there's already immense selection pressure on bacteria to evolve some resistance to antibiotics, and while this mostly has really negative implications, it has one positive implication, which is that, if there's an easy way to do it, it's likely that it'll happen naturally first and not through a surprise attack by a deliberate bad actor.

Then another reason that we're worried about viruses to a greater extent than bacteria is because of their higher transmissibility and the greater difficulty we have disinfecting things from viral pathogens. So, I don't think that antibiotic resistance will be a big priority from the far-future biosecurity perspective. I think it's possible that we're completely wrong about this. I'm very open to that possibility, and what I'm saying is pretty low confidence right now.

Question: Great. Next question. To what extent do small and large scale bio-risks look the same and to what extent do the counter measures for those small and large scale risks look the same, such that you can collaborate with people who have been more in the traditional focus area of the smaller scale risks?

Claire Zabel: That's an interesting question. I think it's a complicated one and a simple answer won't answer it very well. When I think about the large scale risks, they look pretty different for the most part from conventional risks, mostly because they're highly engineered. They're optimized for destructiveness. They're not natural. They're not something we're very familiar with, so that makes them unlikely to be things that we have prepared responses to. They're likely to be singularly able to overwhelm healthcare systems, even in developed countries, which is not something that we have much experience with.

But the second part of the question about the degree to which efforts to address small scale risks help with big scale risks and vice versa, I think that that's somewhat of an open question for us and as we move towards prioritizing in the space, we'll have a better view. There's some actions that we can take. For example, advocacy to get the government to take biosecurity more seriously might help equally with both. On the other hand, I think developing specific counter measures, if we move forward with that, will be more likely to only help with large scale risks and be less useful with small scale risks, although there are counter examples that I'm thinking of right now, so that's definitely not an accurate blanket statement.

Question: When you think about these sort of engineered attacks that could create the largest scale risk, it seems like one thing that has sort of been on the side of good, at least for now, is that it does take quite a bit of capital to spin up a lab and do this kind of bioengineering. But, as you mentioned, stuff is becoming cheaper. It's becoming more widely available. How do you see that curve evolving over time? Right now, how much capital do you think it takes to put a lab in place and start to do this kind of bad work if you wanted to and how does that look five, ten, twenty years out?

Claire Zabel: I don't think I want to say how much it takes right now, or exactly what I think it will take in the future. I think the costs are falling pretty quickly. It depends on what ends up being necessary, so for example, the cost of DNA synthesis is falling really rapidly. It might be the case that that part is extremely cheap, but actually experimenting with a certain pathogen that you think might have destructive capability – for example, testing it on animals – might remain very expensive, and it doesn't seem like the costs of that part of a potential destructive attack are falling nearly as quickly.

Overall, I think costs will continue to fall but I would guess that the falling plateaus sometime in the next few decades.

Question: Interesting. Does biological enhancement fall within your project at all? Have you spent time considering, for example, enhancing humans or working on gene editing on humans and how that might be either beneficial or potentially destabilizing in its own way?

Claire Zabel: That's not something that we've really considered a part of our biosecurity program.

Question: Fair enough. How interested is Open Philanthropy Project in funding junior researchers in biosecurity or biodefense? And relatedly, which would you say is more valuable right now? Are you looking more for people who have kind of a high level strategic capability or those who are more in the weeds, as it were, of wet synthetic biology?

Claire Zabel: Yeah. I think that right now we'd be excited about EAs that are interested in either, potentially, depending on their goals in this field, the extent of the value alignment, and their dedication and particular talents. I think both are useful. I expect that the kind of specialization, for example, either in policy or in biomedical science will possibly be more helpful in the long term. I'm hoping that we'll gain a lot of ground on the strategic high level aspects of it in the next few years, but right now I think both are sorely needed.

Question: Next question. For someone whose education and skills have been focused on machine learning, how readily can such a person contribute to the type of work that you're doing and what would that look like if they wanted to get involved?

Claire Zabel: I don't know. I've never seen anyone try. I think that it would be possible because I think that there's a lot of possibility of someone who has no special background in this area, in general, becoming really productive and helpful within a relatively short time scale and I don't see machine learning background as putting anyone at a particular disadvantage. Probably it would put you at somewhat of an advantage, although I'm not sure how. I think that right now, the best way to go would probably be just to get a Masters or PhD in a related field and then try to move into one of the relevant organizations, or try to directly work at one of the relevant organizations like our biggest grantee in biosecurity, the Center for Health Security. And for that, I think that probably having a background in machine learning would be neither a strong drawback nor a huge benefit.

Question: That's about all the time that we have for now, unfortunately.

Animal Welfare

By Jess Whittlestone, 2017

Introduction

One of the greatest problems in the world today may be the suffering of animals in the meat industry. Roughly 11 billion animals are raised and slaughtered in factory farms each year in the US alone,[1] in conditions likely to cause extreme suffering.[2]

This problem seems to be incredibly neglected. Many experts now believe that animals have conscious experiences,[3] and are capable of experiencing pain.[4] We tend to give much more weight to the suffering of humans than to the suffering of animals – this is potentially a form of "speciesism", valuing animals much less than they deserve.

There are things we can do to help solve this problem. Three main types of intervention look promising: persuading people to change their diets, lobbying for better welfare standards, and developing alternatives to animal products. However, the evidence in this area is not as strong as it is for global health interventions.

This profile sets out why you might want to work on improving animal welfare – and why you might not. This area looks particularly high-impact if you think animals' capacity to suffer is similar to that of humans, and the treatment of animals is unlikely to improve naturally as humanity makes progress.

The case for animal welfare as an important cause area

How can you tell where your resources will do the most good?

[1] "Each year in the United States, approximately 11 billion animals are raised and killed for meat, eggs, and milk." Humane Society of the United States, "An HSUS Report: The Welfare of Animals in the Meat, Egg, and Dairy Industries"

[2] "These farm animals—sentient, complex, and capable of feeling pain and frustration, joy and excitement—are viewed by industrialized agriculture as commodities and suffer myriad assaults to their physical, mental, and emotional well-being, typically denied the ability to engage in their species-specific behavioral needs. Despite the routine abuses they endure, no federal law protects animals from cruelty on the farm, and the majority of states exempt customary agricultural practices—no matter how abusive—from the scope of their animal cruelty statutes. The treatment of farm animals and the conditions in which they are raised, transported, and slaughter within industrialized agriculture are incompatible with providing adequate levels of welfare." Humane Society of the United States, "An HSUS Report: The Welfare of Animals in the Meat, Egg, and Dairy Industries"

[3] In 2012, a group of neuroscientists signed the Cambridge Declaration on Consciousness, asserting that "The weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Non-human animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates."

[4] A recent report from Luke Muehlhauser at the Open Philanthropy Project on "Consciousness and Moral Patienthood" reviews in great detail different ways of assessing the consciousness of different creatures. The report concludes that we still know very little about what features are necessary, sufficient, or indicative of consciousness, and takes a more skeptical view of many of the studies of animal consciousness (pointing out, for example, that there is likely a selection bias in those researchers who choose to study animal consciousness.) However, the report still ultimately suggests assigning relatively high probabilities to the likelihood of consciousness in various animals - 90% for chimpanzees, 80% for cows and chickens, and 70% for rainbow trout - based on how similar these creatures are to humans in various relevant ways. Though these probabilities might seem surprisingly low, they still seem high enough to justify taking the possibility of animal suffering very seriously, especially given the scale of the problem.

- *Heuristics*: We can use rules of thumb to focus our attention. In particular, we might look for *important* problems that are being *neglected* by others, or interventions for which there is lots of evidence, and a possibility to gather more (*value of information*).
- *Quantitative estimates*: We might look at studies which estimate the cost-effectiveness of interventions, based on empirical data from *randomised control trials (RCTs)*.

Animal suffering is large in scale

Each year, tens of billions of animals are raised for meat and slaughtered on factory farms. This is many times more than the total number of humans alive today (~7 billion). A report on the conditions of modern factory farming by Vegan Outreach details how dire these conditions can be. Birds raised for chicken are severely food-restricted in densely populated sheds with large amounts of waste accumulating.[5] Egg-laying hens are packed together in small cages. Many die of asphyxiation and dehydration[6] while male chicks are ground up alive or gassed.[7] Some dairy cows are kept inside all year round[8] and their calves separated from them within 12 hours of birth.[9] Pigs raised for meat are kept in stalls for years where they cannot even turn around.[10]

It is also becoming clear that factory farming is harming the environment. A report by the Worldwatch Institute estimates that animal agriculture accounts for over half of all human-caused greenhouse gases.

The problem of animal suffering looks even larger in scale when we also consider wild animals, which vastly exceed factory farmed animals in number. For example, it's estimated that between 1 and 3 trillion wild fish are caught and slaughtered each year for human consumption.[11]

Animal welfare is neglected

Especially given the scale of the problem, animal welfare seems incredibly neglected. Around 97% of philanthropic funding in the US goes towards helping humans.[12] The remaining 3% is split between the environment and animals. Even within the funding spent on animal welfare, only 1% goes towards farmed animals, although over 99% of domesticated animals are farmed animals.[13] [14] In total, an estimated $10-

[5] Inma Estevez, "Ammonia and Poultry Welfare," Poultry Perspectives (MD Cooperative Extension), 2002; 4(1)

[6] United Egg Producers, Animal Husbandry Guidelines for U.S. Egg Laying Flocks 2016 Edition.

[7] G. John Benson, DVM, MS, and Bernard E. Rollin, PhD, eds., The Well-Being of Farm Animals: Challenges and Solutions (Blackwell Publishing, 2004).

[8] U.S. EPA, Ag 101: Dairy Production, 27 June 2012

[9] USDA APHIS VS NAHMS, Dairy 2014: Dairy Cattle Management Practices in the United States, February 2016.

[10] Bernard E. Rollin, PhD, Farm Animal Welfare (Iowa State University Press, 2003).

[11] "Previous studies suggest that fish experience pain and fear and that, for commercially-caught fish, the severity and duration are likely to be high. This study seeks to assess the number of such animals... Fisheries capture statistics (tonnage by species) published by the FAO were used, along with estimates of mean weights for different species, to estimate the global number of fish caught annually." Estimating the Number of Fish Caught in Global Fishing Each Year.

[12] Source: Charity Navigator's Giving Statistics, 2015

[13] Source: Animal Charity Evaluators, "Why farmed animals?"

[14] "Animal welfare is an area that receives a lot of attention and funding, but most of this is focused on pets and animals used for lab testing; farm animals receive less attention. This may be because people don't want to think about where their food comes from, because they find it disgusting or don't want to feel as if they should change their diet. Certain events have created short-lived increases in public interest, such as Mad Cow disease or the horsemeat scandal. However, this interest wanes, because people generally trust the federal government to address public health issues." Notes from a conversation between the Open

40 million is spent each year on reducing animal suffering in factory farms,[15] roughly a tenth of a cent per animal.[16]

It seems like we can make progress

There is a small but growing base of evidence on animal welfare interventions, suggesting promising ways to make progress on this problem, and that some approaches could be cost-effective.

Campaigns to try to get large companies to reduce their impact on animal suffering are one of the most promising types of intervention. Corporate campaigns to date have resulted in cage-free pledges from around 100 companies, sparing about 60 million hens annually from confinement.[17] Lewis Bollard (Program Manager for animal welfare at the Open Philanthropy Project, and fund manager for the animal welfare Effective Altruism Fund) conservatively estimates that these campaigns will spare about 38 hens a year from cage confinement per dollar spent - and may save up to 250 hens a year from confinement (depending on how the money spent on these campaigns is counted).[18]

Persuading individuals to go vegan/vegetarian may also be a promising approach. The expected value of going vegetarian seems to be significant.[19] This case is strengthened by the fact that each individual going vegetarian may make it more likely that others go vegetarian. Distributing leaflets is cheap, so leaflet distribution could be cost-effective even if only one in every few hundred people decided to go vegetarian as a result. There have been some preliminary studies looking at the impact of leafleting, but the results are not yet particularly conclusive. ACE's leafleting study found a small proportion of people who received a leaflet about reducing animal product consumption did report stopping eating meat, relative to no change in a group who received a 'control' leaflet. However, the sample was too small for these results to

Philanthropy Project and Adam Sheingate.

[15] The Open Philanthropy Project's report on the Treatment of Animals in Industrial Agriculture cites the following main sources of funding in this space, on which this estimate is based: The Humane Society of the United States has an annual budget of $1 million/year for Farm Animal Protection (though resources from other parts of the organisation may also be directed towards farm animals, potentially raising the total up to ~$10 million); six smaller advocacy organisations have annual budgets in the range of $500,000 to $2 million, and Farm Sanctuary, which has a budget of ~$9 million per year.

[16] There are also psychological reasons to expect we might neglect this problem. We tend to be most motivated to help when we can empathise with a specific living being, and when we feel personally responsible for alleviating their suffering. But these motivating factors are absent for factory farming. The problem is somewhat hidden, affects many anonymous beings, and is something no single person feels much responsibility for.
In addition, we may be biased against the suffering of animals.This bias is sometimes called "speciesism". There is experimental evidence that people think of animals as having far fewer mental capacities than they in fact do. This bias is perhaps because of the widespread acceptance of meat consumption and farming practices. For someone who eats meat, acknowledging the harm this causes is likely to be incredibly emotionally difficult. This means that they may have a strong incentive to turn a blind eye and find ways to convince themselves that the problem is not so big after all.

[17] Source: Open Philanthropy Project

[18] "Counting just the ~$2.5 million spent on corporate cage-free campaigning over the last few years, and conservatively assuming that the campaigns only accelerated pledges by five years, these campaigns will spare about 250 hens a year of cage confinement per dollar spent. And even if you add the $1.5 million disbursement for the first year of our three grants, and the ~$12.5 million (at most) spent both on Prop 2 in 2008 and all egg undercover investigations ever done in the U.S., these campaigns will still spare about 38 hens a year of cage confinement per dollar spent. In my view, the assumption that these campaigns only accelerated pledges by five years is very conservative. It seems equally likely that these companies would never have dropped battery cages, or would have merely transitioned to "enriched" cages. For instance, as recently as March 2015, a coalition backed by McDonald's, General Mills, and other major food companies issued a report which largely endorsed "enriched" cages as an alternative to cage-free systems." (From OPP's profile on The Humane League and corporate campaigns.)

[19] Brian Tomasik makes a convincing case that an individual's choice to go vegetarian does make a difference, at least in expectation. It seems that each individual vegetarian is unlikely to make a difference to the number of animals raised. But if they do make a difference, that difference is likely to be very large, so the expected value will be large.

be statistically significant.[20]

Overall, RCTs and cost-effectiveness analyses suggest that there are opportunities to have a large impact on animal welfare. The evidence from corporate campaigns seems particularly promising. However, the evidence base here is still relatively small, which suggests we should be less confident in these estimates of direct cost-effectiveness. This is because we might think that most interventions are not very cost-effective, so we should be skeptical of weak evidence of high impact. However, it also suggests that it may be worthwhile to invest more resources into evaluating the impact of animal welfare interventions. This could help us to make better decisions in future – focusing our efforts elsewhere if these interventions do not seem to be effective, and scaling them up if the results are more promising.[21]

To summarise, we believe that it's possible to have a large impact in animal welfare because:

- *Scale*: Tens of billions of animals suffer and are slaughtered in factory farms annually. Orders of magnitude more animals suffer in the wild.
- *Neglected*: Less than 3% of all philanthropic funding goes towards helping animals, and only a tiny proportion of that funding actually going towards the animals who suffer the most.
- *Cost-effective*: Corporate campaigns have had some large successes in persuading big companies to change their practices, and rough calculations suggest that these campaigns could be extremely cost-effective.
- *High value of information*: Doing more research into this issue could help us to decide how many resources we should devote to the problem.

Some concerns about prioritising animal welfare as a cause area

Some people object to the idea of reducing factory farming by claiming that eating meat is natural – or that we need to eat meat in order to be healthy. There are a few important points to note in response to these concerns.

Eating meat is natural

First, just because something is "natural" does not necessarily mean that it is good. For example, it is natural for children to go unvaccinated, with a large proportion dying at a young age. But this state of affairs seems clearly wrong.

Second, even if some amount of predation were natural and necessary, factory farming is not particularly natural. Hens are not naturally kept in tiny cages indoors, and cows are not naturally separated from their calves. At best, wanting to do things that are "natural" could justify personally hunting animals and eating them – but not buying factory farmed, prepared meat from the supermarket.

Moreover, many of the unnatural conditions in factory farms are avoidable. We have the resources to raise

[20] Source: Animal Charity Evaluators

[21] This is also relevant when it comes to the problem of wild animal suffering, which we mentioned may be many of orders of magnitude larger a problem than even factory farms. Though it's not currently clear how tractable this problem is, the potential gains from exploring potential ways to solve this problem could be extremely high, given the scale and neglectedness of the problem.

animals more humanely, alleviating their suffering.

We need to eat meat to be healthy

Many people defend the mass production of meat saying that we need meat in order to be healthy However, it's far from clear that humans today need to eat animals or animal products, with some evidence suggesting that vegetarians and vegans are in fact more healthy than meat-eaters.[22] [23]

There are still risks that a vegan diet can lead to deficiencies in certain micronutrients, such as B12 and Omega 3.[24] It's relatively easy to find vegan foods that contain these nutrients (such as dark green vegetables and fortified cereals/drinks), and/or to take supplements, but this does require a bit of thought and effort.

Why might you not choose to prioritise animal welfare as a cause area?

Animal welfare seems to be a promising cause area. But there are also a number of reasons why you might be unconvinced by this analysis, or why you might think that a different cause area is likely to hold even greater opportunities to do good.

You might think the evidence for the effectiveness of interventions in this space is not strong enough

The evidence base here is still relatively weak, especially when compared to global health interventions. It is normally clear exactly where and how more money directed towards global poverty can improve and save lives. By contrast, much less research has been done on animal welfare interventions. However, even if there is not enough evidence to say that the intervention is definitely high impact, there seems to be a significant probability that it is very high impact, which could make the expected value high. Some sorts of work, particularly additional research, may also help us to gain more evidence. This is valuable because it helps us make better funding decisions in the future.

You might think we should prioritise helping humans over helping animals

Strength of evidence aside, you might choose to prioritise human-centred cause areas over animals simply

[22] The Oxford Vegetarian Study looked at the health of 6,000 vegetarians and 5,000 nonvegetarian control subjects in the UK between 1980 and 1984. "Cross-sectional analyses of study data showed that vegans had lower total- and LDL-cholesterol concentrations than did meat eaters; vegetarians and fish eaters had intermediate and similar values… After adjusting for smoking, body mass index, and social class, death rates were lower in non-meat-eaters than in meat eaters for each of the mortality endpoints studied… the health of vegetarians in this study is generally good and compares favorably with that of the nonvegetarian control subjects."

[23] "Research shows that plant-based diets are cost-effective, low-risk interventions that may lower body mass index, blood pressure, HbA1C, and cholesterol levels." Nutritional Update for Physicians: Plant-Based Diets.

[24] "Vegans tend to be thinner, have lower serum cholesterol, and lower blood pressure, reducing their risk of heart disease. However, eliminating all animal products from the diet increases the risk of certain nutritional deficiencies. Micronutrients of special concern for the vegan include vitamins B-12 and D, calcium, and long-chain n–3 (omega-3) fatty acids. Unless vegans regularly consume foods that are fortified with these nutrients, appropriate supplements should be consumed. In some cases, iron and zinc status of vegans may also be of concern because of the limited bioavailability of these minerals." Health effects of vegan diets, The American Journal of Clinical Nutrition.

because you think that improving human lives is higher priority. Deciding how to prioritise animal welfare relative to the problems faced by humans depends on a number of complex issues:

1. The significance of animal suffering relative to human suffering

Though it seems likely that animals have the capacity to suffer and feel pain,[25] it seems reasonable to be more confident of human consciousness than of animal consciousness. We have direct evidence of the former, but we still know very little about which physical or functional characteristics are needed to create conscious experience. Depending on how much more confident we think we should be of human consciousness, this might suggest assigning considerably less moral weight to animals than humans, and therefore prioritising interventions in human welfare.[26]

You might also think that there are other reasons that, for example, make humans living in poverty worse than animals being kept in a cage. You might believe that humans' greater cognitive complexity means that their capacity to suffer is greater or more significant. Perhaps freedom and dignity are more important for humans than they are for animals, for example. Or perhaps you think that more complex forms of consciousness – the ability to reason and reflect, for example – are more important than pleasure and pain, and that only humans are capable of experiencing these.

2. The indirect effects of poverty interventions versus animal interventions

Human societies are capable of development in a way that animal societies are not, and so we might think that the indirect effects of human-focused interventions will be greater.[27] However, improving attitudes towards animals might increase our circle of empathy generally, which could itself have positive indirect effects.[28]

However, improving the lives of humans could also have negative indirect effects for animals,[29] since people generally eat more meat as they get richer. This is sometimes known as the meat-eater problem.

3. How likely you think it is that the problems faced by each will get solved anyway

Relatedly, it's also worth considering how likely the problems faced by humans and animals are to be solved without us intervening. There may be good reason to expect that humans will be naturally

[25] Source

[26] It's worth noting that there are a number of complex steps and assumptions involved in reaching a conclusion about whether to prioritise human or animal welfare interventions, especially if we're considering specific types of interventions. First, there is the question of how confident we should be that different types of animals are conscious, which turns out to be incredibly complicated. Next, there is a question of how this assessment of consciousness should be combined with other factors to assess the relative moral weight of humans vs animals - a separate issue, which might depend on things like more detailed theories of wellbeing or how "unified" the conscious experience of different beings are. Finally, even if we can decide how much relative moral weight we should assign to animals, we then need to consider the scale of the problems faced by / the suffering inflicted on different species - even if we assign humans twice the moral weight of animals, for example, we might still prioritise animal welfare interventions if we think that animal suffering is more than twice the scale of human suffering.

[27] We might expect lifting people out of poverty, for example, to have all kinds of knock-on benefits resulting from these people then being able to contribute more to societal progress. Since animals do not live in such organised societies, it seems less likely that there would be such additional benefits beyond just reducing their suffering - more content animals seem less able to contribute to societal progress. Owen Cotton-Barratt suggests that this might be a reason to favour interventions focused on human welfare over those focused on animal welfare.

[28] Though it's worth noting that if your goal is to increase empathy broadly, there may be other / more direct ways to do this than by focusing on animal welfare worth considering. There are also some arguments against explicitly working on moral advocacy here.

[29] Especially if human progress is not accompanied by similar levels of moral progress

motivated to improve their lives and the lives of others. It's less clear whether humans will be naturally inclined to improve the lives of animals.

This might be a point in favour of prioritising them in our altruistic efforts. Even if we think that humanity's ability to empathise with animals is likely to increase naturally, it's not clear this will happen fast enough to outweigh our increasing ability to 'accidentally' harm animals as a result of pursuing other goals. We probably harm many more animals today through the practice of factory farming than our ancestors did hundreds of years ago, despite the fact that our empathy for animals appears to have increased - simply because it's easier for us to harm animals at scale, and to participate in harming animals (e.g. by purchasing meat) without having to cause the harm ourselves directly.

You might think it's possible to have a greater impact focusing on the long-run future of humanity

The long-run future could be incredibly morally important. We could have billions or even trillions of descendants. This means that even a low probability of improving the lives of future generations, or ensuring that humanity survives into the future at all, could be significant.

In the case of animals, focusing more on the long run future might mean focusing more on changing attitudes and reducing speciesism, and less on short-term farming practices.

However, the extent to which you might choose to focus your altruistic efforts / donations on the long-run future, and what that means in practice, depends on some judgement calls:

1. How much do future lives matter?

Since future lives don't exist yet (by definition), some people think we have less moral reason to help them than those who do exist. On "person-affecting" views, an action can only be good or bad if it is good or bad for *someone* - and so actions that will benefit future generations of people or animals cannot be either good or bad, since there is no-one specific they are good or bad for.

If you don't think future lives matter, then you'll want to focus on alleviating the most immediate forms of suffering. It's also worth noting that person-affecting views might also affect how important you think certain kinds of animal advocacy interventions are - since animals live relatively short lives, the impact of many of these interventions will be on animals that don't yet exist, giving less moral weight to these views.

However, we do think there are some convincing reasons to object to person-affecting views (not least that they lead to some counter-intuitive conclusions), and that future beings do matter morally, which we discuss in more detail in our profile on the long-run future.

2. Is there anything we can do to improve the lives of future sentient beings with any certainty?

There is an empirical question of how easy it is for our actions today to have any impact on people and animals living in the long-run future. You might think that future and presently-existing living beings are equally important, but that it's much easier for us to do things that help those currently alive, and that we can therefore have a greater impact by focusing on them.

3. How concerned should we be about the possibility of extinction, and is there anything we can do about it?

If you believe that there is a serious threat of human extinction in the next century (or less), and that there are actions we can take now to reduce that threat, then it could be more important to focus on reducing those threats than on improving the welfare of either humans or animals in the more short-term. Suppose there was a >50% chance of an asteroid hitting the earth in the next 50 years that would wipe out all life. Suppose further we were confident that there were things we could do to detect and prevent the asteroid impact, but that doing so would take a great deal of resources. In this situation, we would probably be willing to prioritise putting resources towards this over putting resources into reducing factory farming right now – precisely because it's little use improving the lives of animals if the whole planet might soon be wiped out.

Of course, this depends a great deal on how likely you think the asteroid (or other existential threat) is, and how confident you are that any attempts to thwart it can make a difference. Short of being 100% certain of the threat and exactly how to prevent it, we wouldn't want to divert all the world's resources away from more immediate problems – but we would likely seriously reprioritise. Even if the threat is relatively small, it seems like we should be investing some resources into anticipating and preventing a potential global catastrophe, given how much is at stake. We think this kind of work may currently be too neglected given its importance, because it is so abstract. This perspective and the reasons to take it seriously are outlined in more detail in our profile on the long-run future.

Summary

- The problem of animal suffering is huge in scale: each year, at least tens of billions of animals live in dire conditions and are slaughtered in factory farms.
- The current state of evidence suggests we should put a high probability on animals being conscious and able to suffer.
- The problem of animal suffering is incredibly neglected relative to its scale, with less than 3% of total philanthropic funding going towards animal welfare, and less than 1% of that funding going towards factory farming, the practice that accounts for more than 99% of animal suffering in the world today.
- There is also good reason to think that the problem is relatively tractable: campaigning of various forms has been shown in the past to be effective at changing individual and institutional behaviour as well as regulations around the treatment of animals, and innovation in the food industry seems a very promising way to make it easier for people to avoid animal products.
- However, these interventions are not as well-proven as some focusing on human welfare, such as global health interventions.
- Whether you believe this to be the most important cause depends on: how important you think it is to have strong evidence of effectiveness; how confident you are that animals are conscious; whether you think our treatment of animals will likely improve anyway, and how important you think it is to focus on the long-run future of civilisation over more immediate problems.

Effective Altruism in Government

By Jason Matheny, 2017

In this 2017 talk, IARPA director Jason Matheny talks about how effective altruists can have an impact through entering government. He draws widely from his own experience in the US government.

I don't have any slides just because they're much harder to get through government review, which will be a big part of my talk today on what to do and not to do in government. It's great to see so many young people who are thinking about career options. I feel like I'm close to dying now, so there's only so much good remaining I can do, unless I'm cryonically preserved and resuscitated later. You all have so many decades left to do an enormous amount of good with your lives, so thank you for spending at least 50 minutes of that life to think about options in government.

I'm going to describe a few different paths to doing good within government that don't rely on spending an entire career in government. It might be a two-year stint, or a five year stint, but still allowing you to accomplish a huge amount of good.

I'm going to talk about three things. The first is how I came into government. The second is the roles for EAs working within government. The third is some practical advice on picking jobs within government, whether short-term or longer-term.

My own route was sort of circuitous in that I started working in college, planning to become an architect. Shortly after I graduated, I found an orphan copy of the 1993 World Bank World Development Report, which both dates me and dates a lot of my perspective on problems of altruism. The report was really focused on how you can do cost-effectiveness analysis on health and development. For me, this was the first time I was ever exposed to an argument about how cheaply you can save lives and significantly reduce suffering. This was one of the first reports that looked at the cost per DALY averted for a range of different health interventions. I thought this application of cost-effectiveness analysis to health and development was pretty eye-opening. For one, it showed that you could save millions of live for less than about $1,000 per life. But it also showed that there were tremendous differences between the most and the least cost-effective interventions, so that our decisions about what to invest in have a huge impact.

I ended up deciding not become an architect, or else I'd be giving a talk on effective altruism in architecture. I decided to instead go into global health, and worked for several years on global infectious diseases – especially malaria, tuberculosis, and HIV in South Asia, but also in East Asia. It was in 2002 that I made another career shift. This was the same year that the first virus was synthesized from scratch, de novo. It turned biology into something more closely approximating computer science, or an engineering discipline. You could now take the chemical constituents of DNA and create a new genome.

At that time, the technology was fairly primitive. The longest virus that you could assemble was polio. That wasn't too much to worry about. We already had vaccines for polio. We knew how to control polio. But I and most of the people who I had worked with had worried that somebody would apply this technology to recreate smallpox – to recreate the 1918 influenza that killed over 50 million people in one year. Or, they could make something much worse than any naturally occurring virus, since there are limits to nature's ingenuity that might be outpaced by human ingenuity.

So I moved from working on naturally occurring global infectious diseases to working on defences against engineered threats. That work then led me to places like the Future of Humanity Institute, thinking about how we wrestle with risks from emerging technologies. I thought about ways that I could have an impact on this area. One way seemed to be doing research myself, but I didn't think I was especially smart in doing that research. It seemed like there were people who were smarter than me, like Nick Bostrom, who could be doing that research. I thought: how could I deliver more funding to people who were smarter than me? That's why I joined government.

I came to IARPA to put a multiplier effect on my effort and the effort of other researchers. My goal was to set up a budget in which I could fund important research in a range of areas including risk assessment, technical forecasting, work on biosecurity, nuclear security, cyber security, and assessments of future risks from things like autonomous weapons or AI accidents. The work that I've done at IARPA has convinced me that there's a lot of low-hanging fruit within government positions that we should be picking as effective altruists. There are many different roles that effective altruists can have within government organizations.

I'm going to tell you about a few of them. I'll limit it to the three areas that I have some background in, which are global health, animal welfare and catastrophic risks. But I think that government is in general a place that has leverage over a variety of different pressing societal issues. Thus if I'm leaving out some topics that's not because I think government doesn't have influence on those topics. It's just because I'm ignorant.

In global health, government funding has an enormous impact on the development of new vaccines, antivirals, and antibiotics, as well as other therapies. Much of that work is conducted at the National

Institute for Allergies and Infectious Diseases, that funds basic and applied research in infectious diseases, as well as the development of new therapies. The FDA has an important role determining which kinds of drugs and vaccines are ultimately introduced. Despite being a regulatory agency, they make really interesting innovations, incentive systems, and reward systems, including things like priority review vouchers that can accelerate the introduction of new vaccines for important diseases. And there are places like the Fogarty Center at the National Institutes of Health, which does things like economic analysis of international health interventions. Then there are the more operational arms of the US government, like the Centers for Disease Control and the US Agency for International Development, that have direct impacts on health and development overseas.

In animal welfare, the USDA has an important role to play in establishing policies that govern some treatments of animals. They also have a small research budget, some of which is used - and a larger amount of which could be used - to develop better, healthier alternatives. State legislatures also have a big role to play in animal welfare as they can pass laws surrounding animal handling and slaughter.

Then on catastrophic risks where I've spent most of my time, governments have a very significant role. A lot of those risks are influenced by government decisions, both positively and negatively. Multiple organizations work on preventing nuclear war, biological warfare, or accidents, as well as cyber warfare, and the misuse of various emerging technologies. I'll go through a few of the most important of those organizations. First you have the National Security Council, a part of the White House that informs national security decision making, including decisions about war. Another is the White House Office of Science and Technology Policy, which looks at emerging technologies and risks associated with them. That office has groups that examine a range of important technologies, including AI, bio-technology, and neuroscience among other topics.

Within the Department of Defense, there is the Office of Net Assessment, which in my view is one of the most unusual organizations within government, as well as one of the most important. It looks at long-range security issues that could be decades in the making. For example, what changes in future weapon systems are likely to disrupt deterrents? What would be the consequences of strategic miscalculation with nuclear weapons?

The Defense Threat Reduction Agency within the Department of Defense is the lead agency responsible for countering chemical, biological, radiological, and nuclear weapons. The Federal Emergency Management Agency is responsible for considering worst-case scenarios that could affect the US and developing mitigations against them. Then there are many others that deal with very specific threats such as the Defense Department's Strategic Command, which is responsible for the United States's nuclear weapons and their safety. BARDA, which is a part of Health and Human Services, is responsible for developing medical countermeasures against bioterrorism. The intelligence agencies like CIA and DIA, assess how advanced a particular group's biological weapons program is, or their ability to access disruptive technologies, or the likelihood of industrial accidents, for example in foreign biology labs.

Across all three of the EA topics that I mentioned, in global health, in animal welfare, and in catastrophic risks, one cost-effective route to having an impact is to affect the funding of new technologies that could in some ways obviate the need for certain kinds of harmful technologies or reduce the risks of technologies by making sure they're sufficiently protected through safety engineering. For funding scientific and technological research, there are a few important organizations within government. There's the White House Office of Management and Budget, which helps to set the White House budget requests. Often we'll find even fairly junior folks who are putting their weight on multi-billion dollar decisions. It really is extraordinary that even fairly junior positions can have incredible influence. If you think of this just in terms of an expected value calculation, even a 10% probability of affecting a $10 billion decision means a billion dollars in expectation, and can be hugely consequential on topics such as nuclear safety, biological safety, future of autonomous weapons and so forth.

There are the Congressional Appropriations Committees [link to senate page] that approve the budgets from the White House. Here, too, you find even fairly junior staffers that have an incredible impact. Then there are the organizations that take the budgets that they've been given and freely decide how to allocate them. Those include places like the National Institutes of Health, the National Science Foundation, as well as the ARPAs - the intelligence IARPA where I work, DARPA, HSARPA and ARPA-E. At those organizations the program managers, who are typically in their 30s say, have come out of graduate school programs and a science or engineering discipline. They've spent a few years working in a lab, sometimes within academia or within industry. Then they spend a term-limited time in the government, usually not exceeding five years. They're given an extraordinary amount of latitude. They're given a budget of several tens of millions of dollars with the expectation, the trust, that they will invest that money as cost-effectively as possible in solving a particular technical problem. For IARPA, those problems are often associated with reducing the risks of emerging technologies.

As one example, we have a brilliant biologist at IARPA, John Julias, who runs a program called Fun GCAT, which is focused on developing new systems for screening the sequences that go into DNA synthesizers. Can you determine whether this is a safe sequence or a dangerous sequence? That's the kind of work that we really need program managers to do, and we've entrusted John with a +$50 million budget which he uses to fund work here at Harvard, at MIT, and at many other universities and companies in order to advance this goal of reducing risks from synthetic biology. It's much more money I think than at least I could've expected to earn in my lifetime, but we give it, we entrust it, to a program manager to spend as wisely as possible - with the rigor of spending a quarter of that money on testing and evaluation to figure out whether the investments that we're making are actually making a difference, and whether we can accurately assess the risks from, say, a novel sequence. Those are program managers.

Agency directors can further direct hundreds of millions or even billions of dollars to key projects. Again, even if those are only, say, 10% as effective as funding that would be given outside of a bureaucracy, the expected value of those investments is quite large and can have a dramatic impact. Within government, there are also other levers that one can pull. At IARPA, we've not only been able to erect a large budget on reducing catastrophic risks, but we've also been able to engage in policy discussions. We've led groups within the White House on the long-term impacts of AI and of biotechnology. We co-led the White House AI R&D strategy, and we've advised the National Security Council on other emerging technologies. There are some decisions that are made only by governments, and some of those decisions are highly consequential. They include decisions like going to war, or what weapon systems will be fielded, or how technologies will be embedded within larger critical systems. It makes sense to engage more effective altruists within these positions where they can influence those decisions.

One can also have influence on the outside, working as a contractor within a government agency. Most of the people who work at IARPA are contractors rather than government employees. The amount of expertise that we have to draw on is too vast to hire them all ourselves directly, especially with short-term positions. So we hire computer scientists, and biologists, and chemists, and physicists, and neuroscientists, and sociologists, political scientists, and cognitive psychologists, because we need them all. We also need lawyers, and we even need philosophers. We have a program on applied philosophy called CREATE, which is a program to develop new systems for argumentation and informal reasoning that can lead intelligence analysts to make better judgements.

So we need lots of help. We need them from lots of places including contractors, but also think-tanks. There are a range of think-tanks that inform the policy-making process that sometimes have a quite deep influence on administration. For instance, the Heritage Foundation has a substantial influence on the current administration, while past administrations have been influenced by other think tanks such as Brookings, the Center for a New American Security, the Center for Strategic and International Studies, and Harvard's own Belfer Center. There are many others that help shape the decision-making of government leaders. Hence, that's another way you can have an influence on government.

I'm going to close in my final few minutes just by providing some general advice if you're interested in pursuing a job, whether short-term or long-term, within government. I would recommend really thinking about opportunities to move across and among the different sectors of society, government, industry, academia, and NGOs, because there's a need for horizontal transfer of knowledge and best practices. If all of our folks within government have come from government straight out of school, that will prevent us from being able to adopt best practices from industry or from academia. So there really is a need I think for continuous cycling throughout a career, bouncing around between the different sectors in order to bring knowledge across them.

My first suggestion is to reach out to 80,000 Hours, which I think has been pulling together some advice about government jobs (see below). I think one of the pieces of advice is at least to consider it as an option, because we are nowhere near the saturation point of effective altruists going into government positions. There are fairly junior positions across government that have a high potential impact that we have trouble recruiting for.

My second suggestion is to get to know the people who work within the organizations where you'd like to work. You can learn a lot about those organizations, their structure, and their staff just from online websites as well as Wikipedia. You can find the biographies of some of the people whose careers you might want to mimic. One strategy is just to reverse engineer their biography. Figure out what the steps are that seem critical in getting to the positions that you would like to have in the future. On that point, many of these people are lawyers, but just as many of them are scientists and engineers, and we do need more philosophers in government as well. But I think you'll find the diversity of talent that we need is ever growing. There's a particular intersection between policy and technology that is extremely difficult to recruit for. So for people who are still picking their major or their concentration for a thesis, if you look at the science policy of blank, pretty much any of the topics that are critical on our list have not been saturated with attention. There's still lots of low-hanging fruit to pick.

My last suggestion is to reach out to me, especially if you're interested in pursuing a job, short-term or long-term, in national security or reducing global catastrophic risks. Mostly because I really need the help. That's it.

Global Health and Development

By Jess Whittlestone, 2017

Introduction

In 2013 nearly 800 million people were living under the international poverty line.[30] This has a huge negative impact on health[31] – each year, millions of these people die from preventable diseases such as malaria, tuberculosis, and diarrhoea.[32]

This immense suffering is easily preventable, but is nevertheless neglected – OECD governments spend on average just 0.32% of their GDP on foreign aid.

This profile sets out why you might want to focus on problems in global health and development – and why you might not. This area looks most promising if you are sceptical of our ability to influence the longer-term future of the world, if you think that animal suffering is not as significant as human wellbeing, or if you think we need strong evidence of impact to justify interventions.

The case for global health and development as an important cause area

How can you tell where your resources will do the most good?

- *Heuristics*: We can use rules of thumb to focus our attention. In particular, we might look for important problems that are being neglected by others, and interventions for which there is lots of evidence, and/or where learning more would be very valuable.
- *Quantitative estimates:* We might look at studies which estimate the cost-effectiveness of interventions, based on empirical data from randomised controlled trials (RCTs).

Global poverty causes a great deal of suffering for a huge number of people

In 2013, 10.7% of the global population lived below the global poverty line of $1.90 per day – that's nearly 800 million people.[33] This line is intended to represent the minimum level of income needed to fulfil basic needs: food, clothing, healthcare and shelter. Arguably the biggest negative impact of poverty is the

[30] From The World Bank's Poverty and Shared Prosperity report in 2016: "In 2013, the year of the latest comprehensive data on global poverty, 767 million people are estimated to have been living below the international poverty line of US$1.90 per person per day. Almost 11 people in every 100 in the world, or 10.7 percent of the global population, were poor by this standard, about 1.7 percentage points down from the global poverty headcount ratio in 2012."

[31] The impact on health is of course not the only negative consequence of poverty, and improving health outcomes is not the only way to combat poverty. However, the most effective poverty interventions we currently know focus on developing world health, so we focus mostly on those interventions in this profile.

[32] According to data collected by the World Health Organisation (WHO), nearly half a million people died of malaria in 2015, diarrhoeal disease kills around half a million children under 5 each year, and 1.8 million died from tuberculosis in 2015 - in all cases a large majority of the deaths occurred in low-income countries. A study published in the Lancet in 2003 also reports that more than 10 million children die each year, mostly from preventable causes and almost all in poor countries.

[33] The global poverty line was first established as living on less than $1/day, and has since been adjusted for inflation.

cost to health:[34] millions of people die each year of diseases such as malaria or tuberculosis that are easily preventable or treatable in the developed world.[35] It is estimated that the damage done by these diseases in the least developed countries plus India is between 200 and 500 million DALYs (disability-adjusted life years, a measure of a year of healthy life) per year.[36]

There are well-evidenced ways of reducing poverty

Although it is a huge problem, global poverty seems to be relatively tractable – especially if we focus on the immediate costs to health and quality of life.[37] In particular, we know how to prevent and treat the most common diseases, and the cures are often relatively cheap. For example:

- malaria can be prevented by giving people insecticide-treated bednets,[38]
- tuberculosis can be treated with a course of drugs over 6 to 9 months,[39]
- diarrhoea can be prevented through better sanitation[40] and treated with oral rehydration therapy,[41]
- parasitic diseases can be cured with a pill that costs under $1.[42]

(We discuss the cost-effectiveness of these kinds of approaches in more detail in the next section.) Over the last 60 years, millions of lives have been saved using these techniques, suggesting a very clear way to make progress.[43]

Additional resources could do a great deal more in this area

Global poverty does receive a lot of attention from individuals and organisations. However, the total funding going towards poverty alleviation is small relative to the funding for many problems in developed countries. The UK government spends roughly 0.7% of national income on foreign aid each year ($12.1

[34] In a 2006 paper, "The Economic Lives of the Poor", development economists Banerjee and Duflo used survey data from 13 countries to document the lives of people living in poverty, providing more evidence of the impact of poverty on weakness and disease. In Udaipur, India (the country which had the best data) they found that 65% of poor people were underweight (having a BMI of below 18.5), 55% had an insufficient red blood cell count (anaemia), 72% had at least one symptom of a disease, and 46% had recently seen a doctor because of a disease.

[35] The WHO's fact sheets explicitly state that malaria is "preventable and curable", using insecticide-treated mosquito nets and indoor residual spraying, and that TB is "treatable and curable" with a standard 6 month course of antimicrobial drugs.

[36] From the 80,000 Hours problem profile on health in poor countries: "The population of these countries is around 2 billion. To prevent 100 million DALYs each year each person in these countries would have to be given an average of 1/20th of a DALY each year. Given an existing life expectancy of around 65, this would require extending life expectancy by 3.25 years, or the equivalent in improved quality of health. This seems possible and if anything small relative to health gains achieved by other countries that have eliminated easily prevented diseases in the past."

[37] This isn't to say that focusing on preventing and treating disease is the only way to tackle poverty-related suffering, but it's the approach that we currently have the best evidence for, where it's clear one can make a big difference for relatively little money.

[38] Source: World Health Organisation (link to Givewell website)

[39] Source: Center for Disease Control and Prevention

[40] Source

[41] Source: World Health Organisation

[42] GiveWell estimate that Deworm the World Initiative can deworm children for around $0.32 per child in India, or $0.79 in Kenya.

[43] Between 2010 and 2015, mortality rates from malaria fell by 29% globally, and an estimated 6.8 million malaria deaths have been averted globally since 2001 (source). An estimated 49 million lives were saved through TB diagnosis and treatment between 2000 and 2015. (source). The annual number of deaths attributable to diarrhoea among children aged under 5 years fell from the estimated 4.6 million in 1980 to about 1.5 million today (source)"

billion in 2015) – but spends nearly three times as much (around 2% of national income) on defence.[44] Most other countries in the world spend a similar or even smaller percentage on foreign aid.[45]

Individual donations to overseas aid are also relatively small. In the UK in 2015, people donated more to each of medical research, hospitals and hospices, religious charities, and charities for children and young people, than they did to overseas aid and disaster relief,[46] and there are similar results from the US.[47] The charities currently implementing the most effective global health interventions seem to have a clear need for more funds – for example, charity evaluator GiveWell estimates that the Against Malaria Foundation could productively use at least $78 million and potentially up to $191 million in additional unrestricted funding this year.

We have stronger evidence for interventions in this area than almost anything else

There is a robust record of success in global health and development, and the relevant outcomes are somewhat measurable. If, for example, you donate to the Against Malaria Foundation, you can be reasonably certain that you are paying for the distribution of bednets that will prevent at-risk people from contracting malaria.

Though there may be larger potential gains in other cause areas, there is also generally less certainty about what the actual benefit will be. There is a tricky tradeoff here – between a more definite impact, and a less certain but potentially larger impact – which we will discuss in more detail later.

Together, the above analysis provides a compelling initial case for global health and development as an important cause area: it is large in scale, apparently tractable and relatively neglected, and is backed up by strong evidence.

Cost-effectiveness analyses and RCTs

Cost-effectiveness analyses attempt to quantify how much good can be done with a given amount of money.[48] Although these analyses require us to make a number of simplifying assumptions, we believe they form an important part of assessing the impact that can be had in a given cause area.[49]

It is possible to save lives via health interventions in the developing world at very low cost. The

[44] Source

[45] The US government spends less than 1% of the annual budget on development priorities abroad (Centre for Global Development). CGD also note that the public vastly overestimate the percentage of the US budget going to foreign aid - the average answer to a poll taken by the Kaiser Family Foundation was that 28% of the budget was spent on foreign aid.

[46] Source

[47] Charity Navigator's 2015 report on giving statistics in the US found that giving to international charities was less popular than all of the following types of charitable causes: religious causes, education, human services, foundations, health, public society benefit, and the arts. Only 4% of total donations went to international charities.

[48] Or equivalently, how much money it costs to create a specific beneficial outcome - e.g. how much money it costs to save a person's life or give someone a year of additional healthy life.

[49] GiveWell discuss the use of cost-effectiveness estimates to evaluate interventions, and some of the difficulties with doing so, in more detail here.

eradication of smallpox, for example, is estimated to have cost around $1.6 billion.[50] If we conservatively estimate that this saved 60 million lives,[51] this works out at ~$25 per life saved.

This is ridiculously effective when we consider that the UK's National Health Service will spend tens of thousands of pounds to save just a year of healthy life.[52]

The best health interventions currently available may not be quite this effective, but are still extremely promising. GiveWell's latest cost-effectiveness analyses use "saving the life of an individual under 5" as their benchmark, and their estimates suggest that their recommended charities can do the equivalent of this for between $900 and $7,000, depending on the charity.[53] This still seems like an exceptionally good deal compared to the cost of saving lives in the developed world.

These cost-effectiveness estimates rely on some subjective inputs, but they are based on the results of high-quality randomised controlled trials (RCTs). RCTs involve giving a treatment (e.g. insecticide-treated bednets) to half of a population, randomly selected, while the other half are given no treatment (or a placebo/control). We can then measure differences in outcomes between the two groups, where any differences should be due to the treatment alone. These RCTs should therefore give unbiased estimates of the impact of the treatment.

Interventions in global health and development appear to be backed up by more RCTs than we have seen with any other cause area. The distribution of insecticide-treated bednets to prevent malaria, for example, has been studied extensively, and assessed in two Cochrane reviews.[54] Lengeler's (2004) review, which considers more studies and looks at a broader range of outcomes, finds a statistically significant effect on child mortality, summarised as "5.53 deaths averted per 1000 children treated per year."[55]

Another RCT looked at the impact of GiveDirectly's unconditional cash transfers on developing countries.[56] It found that recipients increased the value of their assets, and also saw increases in food security, revenue, psychological well-being, and female empowerment.

To summarise, we believe that it is possible to have a large impact in global health and development because:

[50] The Center for Global Development estimates that the global cost of smallpox - both direct and indirect - in the late 1960s was more than $1.35 billion - and that around $23 million was spent per year between 1967 and 1979 on an intensified eradication programme.

[51] UNICEF estimate that the eradication of smallpox saves 5 million lives annually. If we assumed that without the eradication effort, 5 million lives would have been lost per year to smallpox up to the present day, then this would suggest 190 million lives saved. However, it seems reasonable to assume that smallpox would have been eradicated - or smallpox deaths reduced considerably - at a later date. If we more conservatively assume that smallpox would have continued to kill 1.5 million per year, this suggests ~60 million lives saved up the the present day.

[52] NICE consider interventions that cost the NHS less than £20,000 per QALY (quality-adjusted life year) to be cost-effective.

[53] For example, they estimate that the Against Malaria Foundation can save the life of an individual under 5 for $3,162 by distributing malaria nets, and that Deworm the World can do so for $901. You can learn more about how GiveWell conduct their cost-effectiveness analyses here, and also look in detail at the model they use.

[54] Lengeler, 2004.. Gamble, Ekwaru, and ter Kuile, 2006.

[55] Evidence for the effectiveness of distributing malaria nets is discussed in more detail in GiveWell's intervention report.

[56] Source

- The problem of global poverty is large in scale, affecting hundreds of millions of people and severely reducing their quality of life;
- Many of the problems associated with poverty, particularly the impact it has on health, are highly tractable, since we can treat many of the worst diseases easily and cheaply, and just need to scale up known approaches;
- The problem is relatively neglected given its scale;
- The evidence in this area is particularly strong compared to interventions in other cause areas, and suggests these interventions are cost-effective.

Some concerns about prioritising global health as a cause area

Here we summarise and respond to a number of common concerns about prioritising global health and development as a cause area:[57]

Does foreign aid really work?

A common concern is that developed nations' attempts to help those in poverty are wasted effort and money.

However, this simply does not match the facts: when we look at what aid has achieved over the past sixty years, there's a lot of good to show for it. Though some aid may do little or no good, there's a convincing argument that the average dollar spent on aid has been well worth it. As mentioned above, the lives saved by money spent on the eradication of smallpox work out at ~$25 per life saved, using conservative estimates. Even if we assume that all other aid spending has been completely useless, we could still easily justify the total money spent on the grounds of smallpox eradication alone. If we assume an upper bound of $4 trillion spent on aid, smallpox eradication alone would still give us a figure of around $67,000 per life saved, approximately what the NHS in the UK will spend to save two years of life.

In addition, it's clear that other things beyond smallpox have had a positive effect on poverty. Using nets, indoor spraying, and medicine, we've seen a 29% decrease in annual deaths from malaria between 2010 and 2015.[58] Oral rehydration therapy has cut annual diarrhoeal deaths from 4.6 million in 1980 to about 0.5 million today.[59] Undoubtedly, not all aid works, but this isn't an argument against working to improve health and living standards in the developing world. Instead, it's an argument for demanding higher standards of evidence of effectiveness before we channel significant effort and funds into it.

Charity "begins at home"

Some people think that we should first focus on helping people close to us geographically. Only once problems close to us are resolved should we help the global poor.

However, there are strong reasons to think that additional resources can do a lot more good in the

[57] Many of these are discussed in more detail in Giving What We Can's piece on "Myths about aid."

[58] Source

[59] In 2000 the WHO reported cutting diarrhoeal deaths from 4.6 million in 1980 to 1.5 million, and now reports that diarrhoea kills around 525,000 children under 5.

developing world than they can in richer countries. Simply put, an additional dollar is worth more when you have less money. Precisely because the developing world lacks resources, their biggest problems are ones that we have already figured out how to solve in richer economies, which have a much higher level of health and education. Preventing someone from getting a deadly or debilitating disease improves their life a huge amount, and it's much harder to give someone a similar boost when their basic needs are already met.

In Doing Good Better, William MacAskill suggests that a dollar is worth about 100 times more to someone living in poverty than to the average person in a rich country. This is based on the fact that the annual consumption of someone on a median US income is about 100x that of people living in the most extreme poverty, and the importance of additional money seems to decline with income.[60]

Given how much more effective the same money can be if given to poorer people, it seems hard to justify focusing on those close to us. If Macaskill is right, you would have to think that people close to you are somehow a hundred times more valuable than poorer people overseas. This seems unjustifiable, especially given that many philosophers think that there are no reasons to discriminate against people based on their location or nationality.[61]

Who are we to say what poor people need?

Another concern is that we cannot help people in developing countries without an intimate understanding of their situations and needs.

There are two important points to make in responding to this view. First, global health and development interventions aren't necessarily paternalistic. Recently research has begun to look at the potential benefits of direct and unconditional cash transfers to very poor people, with promising results, and these certainly provide a good baseline against which to compare more targeted interventions.[62] Second, concerns about paternalism seem less well-founded when people have clear needs, such as the prevention of severe malnutrition or suffering. If people are dying of known preventable diseases, it seems hard to imagine how we could be "wrong" about the need to prevent this. There may well be other things that they need beyond surviving, but sorting this out first seems fairly uncontroversial.

Why might you not choose to prioritise this cause area?

Global health and poverty seems to be a promising cause area. But there are also a number of reasons why you might be unconvinced by this analysis, or why you might think that a different cause area is likely to hold even greater opportunities to do good.

You might think that there are better ways to improve the lives of people living

[60] "What the conclusions from the economic studies suggest is that the benefit you would get from having your salary doubled is the same as the benefit an extremely poor Indian farmer earning $220 a year would get from an additional $220. As noted earlier, the typical US wage is $28,000, so there is good theoretical reason for thinking that the same amount of money can be of at least one hundred times as much benefit to the very poorest people in the world as it can be to typical citizens in the West." William Macaskill, Doing Good Better, p.27

[61] Philosopher Peter Singer develops the moral argument in "Famine, Affluence, and Morality".

[62] Source

today

You might believe that focusing on global health and development is simply not the best way to improve the lives of people living today. There might be other problems that people face today which are larger in scale, more neglected, or more clearly tractable – for example, a case could be made that mental health problems create more suffering overall than even poverty does. Or it might be that investing in broader cause areas – such as improving collective decision making, or certain forms of political advocacy – could improve our ability to solve all the problems facing humanity, and so be more effective.

One consideration here is how much importance we should put on having a strong evidence base and track record when prioritising cause areas. We discussed earlier how the strength of evidence is a substantial point in favour of global poverty interventions. However, care is needed to avoid the "streetlight fallacy". We want to look for the best solutions, not just those that are easiest to see or measure. There may be other opportunities to help the world today which have less robust evidence behind them, but have higher expected value because they would do so much more good if successful, and where learning more could be incredibly valuable.

You might think that we should prioritise reducing the suffering of non-human animals

Each year, it is estimated that over 50 billion animals globally live in conditions of extreme suffering before being slaughtered in factory farms.[63] As we discuss in our animal welfare profile, this issue has even less money spent on it than global poverty.

This suggests that animal welfare could plausibly be even larger in scale, and more neglected, than global poverty. Comparing the two turns on the following judgement calls:

1. The significance of animal suffering relative to human suffering

Though it seems likely that animals have the capacity to suffer and feel pain,[64] you might believe that humans' greater cognitive complexity means that their capacity to suffer is greater or more significant. Or you might think that other reasons make it worse that humans live in poverty than that animals are kept in cages. Perhaps freedom and dignity are more important for humans than they are for animals, for example.

2. The indirect effects of poverty interventions versus animal interventions

Human societies are capable of development in a way that animal societies are not, and so we might think that the indirect effects of human-focused interventions will be greater. However, improving attitudes towards animals might increase empathy generally, which could itself have positive indirect effects.[65]

3. The importance of a strong evidence base

[63] Source

[64] Source

[65] Though it's worth noting that if your goal is to increase empathy broadly, there may be other / more direct ways to do this than by focusing on animal welfare.

As mentioned above, global health interventions tend to have much more evidence behind them than animal welfare interventions. If you think that a strong evidence base is important, this might be a reason to prefer global poverty interventions to animal welfare interventions.

You might think that we should prioritise the long-run future of humanity

There could be many more people in the future than are alive today. So if we think we can affect the long-term future, this might be higher impact than focusing on more immediate problems.

This comparison turns on a number of judgement calls:

1. How much moral weight should we give "future people"?

We generally feel an intuitive obligation to treat future, not-yet-existent people in roughly the same way as existent people. However, some philosophers have questioned whether we should give the same degree of moral consideration to future people. On "person-affecting views", an action is only good or bad if it is good or bad for someone – and so the value of an action depends only on how it affects people who are either already alive, or who will come into existence regardless of our actions. One implication of this view is that human extinction, while bad for the people who die, causes no longer-term harms: there is no harm in people failing to come into existence.

A related issue is the non-identity problem, arising from the fact that sometimes future people may owe their very existence to choices made today. For example, which policies a government chooses to enact will affect which people have certain jobs, affecting which people meet and marry, and therefore causing different people to be born in future. Those very policies might also affect how good the lives of future people are – if the government chooses to prioritise policies that increase short-term economic productivity over mitigating climate change in the longer-term, say, this could have a negative impact on future generations. But if the very policies that appear to have made future people's lives "worse off" also ensured that those exact people were born at all, can those people really be said to have been harmed by those policies? If not, then this may be reason to prioritise the welfare of people who already exist, or whose existence does not depend on our actions.

However, the non-identity problem might also be taken as a reason to reject person-affecting views. The implication, that choosing policies that will make future generations lives worse off is not causing those future people any harm, seems highly counterintuitive. We could instead adopt impersonal principles for evaluating the moral value of actions.[66] This means that we would judge not based on how they affect specific people, but based on how good they are from the perspective of the world as a whole.[67]

2. Can our actions today have any real impact on the far future?

[66] This is roughly the perspective taken by philosopher Derek Parfit, who says that it's so obviously good for us to help future generations that this itself gives us a definitive reason to reject person-affecting principles.

[67] Another possibility, which Parfit discusses, is to adopt wider person-affecting principles: these say that while most harms / benefits are comparative (i.e. to say we harmed a person is to say they would have been better off had we acted differently), not all are. In particular, we can say that someone was benefitted by being brought into existence if their life is overall positive. On such views, one state of the world is worse than another, even if different people exist in the two worlds (and therefore neither state is strictly speaking worse for anyone), if the lives of the people in World A are less good for those people than the lives of people in World B are for them.

Even if the future of humanity is incredibly important, you may still believe that there is very little we can do to reliably shape it. The very far future is an area for which we do not have – and cannot have – robust randomised controlled trials and cost-effectiveness estimates.

However, we do think there are reasons to be optimistic here. Small changes in the values of a civilisation could last a very long time, since people tend to try to pass their values onto their descendants. And in the past, relatively small actions have essentially averted global catastrophes: for example, Stanislav Petrov, a lieutenant in the Soviet Air Defence Forces, may have prevented a nuclear war when he judged that reports that the US had launched a nuclear missile were false (which they were).[68]

In addition, we don't need to be highly confident that our actions will have the desired impact, if the potential gains are large enough. Working on the world's largest problems will always be difficult and involve some risk, but that doesn't mean that we should always focus on easier, smaller-scale issues.

3. How should we trade off fixing immediate problems against thinking about the longer-term?

Even if you believe that the long-run future is ultimately what's most important, you might still think that fixing the world's most immediate problems is the best way to influence our trajectory.

If you think that the world is on a positive trajectory, then it makes sense to focus your efforts more on ensuring that humanity survives to enjoy that future. But it might be that there is a serious risk of us getting "stuck" in certain negative patterns, including inequality. In that case, it's possible that one of the best ways to ensure improvements in the longer term is to invest a lot of resources in solving the problems we face right now, to ensure they don't continue to affect future generations.

You might think that we need to focus more on "systemic change"

A final objection to the case outlined above is that it only tackles the symptoms of poverty, not the root causes, and neglects the importance of systemic change. It's not clear whether focusing on the most immediately obvious problems will help us to eradicate poverty altogether. Instead, we may need to better understand the systems in the world that perpetuate poverty and inequality, and think about longer-term strategies for changing these systems.

It is uncertain whether there is a limit to the benefits of focusing on concrete, tractable things like reducing disease, without also focusing efforts on more systemic change. On the one hand, perhaps by helping the worst off in tractable ways, we can get everyone to the level at which they can fulfil their own basic needs, which could naturally lead to economic growth and more productive societies.[69]

On the other hand, it might be that inequality is perpetuated by more fundamental things, such as the politics of developing countries. If this were the case, we might need to address politics more directly.

[68] Source

[69] The basic idea here is that developing countries may be caught in "poverty traps" - self-reinforcing mechanisms that cause poverty to persist, and that things like poor health may contribute considerably to these cycles. If foreign aid can help get people up to a certain level - by improving health and providing opportunities for investment, for example - it may break these self-reinforcing cycles of poverty. Jeffrey Sachs, in The End of Poverty, for example, suggests that: "if the foreign assistance is substantial enough, and lasts long enough, the capital stock rises sufficiently to lift households above subsistence." (Sachs, Jeffrey D. The End of Poverty. Penguin Books, 2006. Pg. 244)

However, it might also be true that while we do need to consider the bigger picture and conduct research to better understand systemic issues, current marginal resources are still more effectively spent directly helping the very poorest people.[70]

We've given some examples in this profile of the kinds of interventions that currently seem most promising for tackling global poverty – but this isn't to suggest these are the only ones worth focusing on. Accounting for less immediate, tangible impacts may be needed to have more of an impact on poverty in the long-run, and things like political action could be very valuable. Evidence for interventions in this space is currently lacking, but more research on opportunities in this area could be very valuable as it enables us to learn more.

Summary

- Global poverty causes a huge amount of suffering across the globe, with over 800 million people affected. One of the worse consequences of this is the widespread existence of diseases which would be easily prevented or treated, and which cause between 200 and 500 million DALYs of harm each year.
- There are a number of interventions which we know to be highly effective in preventing and treating these diseases. These include using bednets to prevent malaria, and distributing treatments to prevent parasitic diseases. There is still plenty of scope for further funding to increase the impact of these interventions further.
- We have stronger evidence for effective interventions in this area than in most other areas.
- In light of this, the objection that foreign aid doesn't work appears to be false. On average, aid does work. It would be even more effective if we focused on the most effective interventions.
- Similarly, since interventions in the developing world can do so much more good than interventions focused on developed countries, it is hard to justify working closer to home.
- It is plausible that we should further systemic change in order to solve these problems. However, systemic change might be best furthered by short-term work which puts people in a position to improve their own futures.
- Whether you believe this to be the most important cause depends on: whether you believe there to be other (perhaps less tested) ways to improve human lives, how much value you give to reducing animal suffering, and whether you believe it to be more urgent to focus on the long-term future of humanity.

[70] Development economist Chris Blattman argues for this kind of perspective, saying that: "We have to focus on the big picture and growth as a society, but I think there's a strong argument for directly tackling the worst poverty now. Especially because we know how to do that pretty well. And we could do it even better pretty easily. More so than figuring out the secret to growth."

How valuable is movement growth?

By Owen Cotton-Barratt, 2015

In this 2015 article, Owen Cotton-Barratt argues that a very important question for social movements is how to grow. He argues that to avoid hostility, it may be wise to prioritize improving people's inclination towards the movement ("advocacy") over making the movement more widely known ("publicity").

Foreword November 2017

I wrote this article two-and-a-half years ago. Although I still agree with its points, I would write a somewhat different article today, in framing and substance.

On the framing, this article uses effective interventions to relieve global poverty as the main example of an idea one might to bring attention to. At the time this seemed the canonical central example for the effective altruism community, who were a good fraction of my audience. Even then I guessed it was not the most important idea to draw attention to, and I have become more confident since. Of course all of the arguments still apply to effective poverty interventions; I just wouldn't lend them so much implicit endorsement.

On the substance, I omitted an important consideration: that larger movements come with increased overhead costs for communication and coordination within the movements. Combined with diminishing returns from effort within an area, this reduces the value of making a movement very large: doubling the size might be substantially less than doubling the value. Moreover if one realises that different people have different amounts to contribute to work on a given problem, the effect is that the value of bringing different types of people in varies significantly. The ideal for a given community might be to include as many as possible of the people who are very well placed to contribute, and not so many people beyond that so as not to impede coordination between that core. For some ideas or problems that might in practice mean quite a small community; for others a large one. The central example of global poverty made this omission less salient since it benefits from more supporters even when large.

Nonetheless, I think the original article contains useful ideas which I continue to make use of in my own thinking. I hope it remains useful to others.

Executive Summary

Movement growth may be very important for young social movements. It's obvious that movement building cannot be always better than direct work, but knowing how to compare them presents a challenge.

In this article I introduce and explore a model of movement growth which tracks individuals' awareness of and inclination towards the movement. I aim to understand when movement building activities are better or worse than direct work, and apply the model to give my views on movement growth in the effective altruist and related communities.

Part 1: Theory

In the first half of this paper I introduce a model for thinking about movement growth, and terminology to refer to critical concepts. We model individuals as having varying levels of awareness about the movement, and varying inclinations towards it. We assume that these two characteristics can represent the major drivers of interaction with the movement. We explore the consequences this Awareness/Inclination Model (AIM), particularly looking at the long-term counterfactual effects of direct work compared to '*publicity*', which aims at increasing awareness of the movement, and 'advocacy', which aims at improving inclinations towards the movement. This involves analysing different possible long-term trajectories the movement may be on.

If we accept the model, this has some general implications:

- For early-stage movements, the effects on movement growth are a key consideration in deciding between different activities. For relatively mature movements, direct work is usually better than movement growth.
- It is more important to focus on increasing awareness than improving inclination, if:
  - the movement has a natural maximum size that we cannot change; or
  - essentially everyone will join the movement after they know enough about it; or
  - direct work earlier is much more important than direct work later; or
  - it is very hard to change inclination relative to awareness.
- Otherwise improving inclination may often be better than increasing awareness (this is sensitive to beliefs about some parameters).
- It is particularly key to avoid being controversial and focus on improving inclination rather than increasing awareness, if:
  - the views of people around them have a significant effect on the inclinations of people towards the movement; or
  - the movement might plateau at a wide range of sizes, depending on how well-perceived it is; or
  - building political consensus will be useful for the direct work.

Part 2: Application

In the second half of the paper, I apply the conceptual tools developed in Part 1 to answer questions about how to find the best work for the young effective altruism movement and related areas. The conclusions here are not certain, but represent my informed best judgement. Some of them are driven purely by qualitative considerations, and some are based in part on numerical estimates.

My conclusions are:

- Getting movement growth right is extremely important for effective altruism. Which activities to pursue should perhaps be governed even more by their effects on movement growth than by their direct effects.
- Increasing awareness of the movement is important, but increasing positive inclination is at least comparably important. Therefore we should generally:

    - prefer *advocacy* to *publicity*;
    - strive to take acts which are seen as good by societal standards as well as for the movement;
    - avoid hostility or needless controversy.
- Direct work has a very important role to play in movement building. It is likely to increase positive inclination, by:
    - Demonstrating commitment, and showing that the people engaged in the movement think the work is valuable;
    - Increasing the credibility of the area by demonstrating that there is productive and valuable direct work that can be done.
- Within global poverty, work focusing on movement growth may be more effective than direct work for most people (at the margin today).
- Within areas that seem promising but do not have an established track record, direct work aimed at demonstrating that there are credible interventions may be one of the most effective forms of movement building.

Acknowledgements: Thanks to Tom Ash, Daniel Dewey, Seb Farquhar, Michelle Hutchinson, Josh Jacobson, Toby Ord, and Kerry Vaughan for helpful conversations and comments.

Introduction

Suppose you have discovered a great way to help the world. Now you face a trade-off: how should you balance pursuing this directly against spreading the word and convincing others to help pursue it?

The basic problem is similar for many possible ways of helping the world. Perhaps you've worked out that washing your hands before surgery helps your patients' survival rate. Perhaps you've realised that treating men and women equally will help welfare levels and productivity. Perhaps you think that giving money to charities that have been shown to do a great deal with it is an effective way to help people. In each case you might have more impact by growing the movement of people who care about the issue than just doing your part of the direct work,[71] but it may be challenging to see how to compare these against each other.

In order to make the best decisions, it's useful to think about how valuable growing the movement is compared to the direct work. This is very hard to know for sure, since we never get to see how all the counterfactuals play out. But I think we can improve our intuitions by understanding some simple models and getting a clearer picture of the kind of dynamics that may be at play.

Section 1, the majority of the paper, aims to give the reader a set of conceptual tools for thinking about movement growth. We introduce the Awareness/Inclination model, and use it to explore the possible life-cycles of movements without intervention, and the counterfactual effects of interventions.

In Section 2 we combine these insights with personal impressions of the effective altruism movement to provide recommendations.

[71] In this paper I count earning to give to direct work as direct work. It is of course indirect in an important way, but in a direct work / movement growth split, it fits with the direct work. In one sense, the idea of 'earning to give' is one example of a way to help the world that a person might have discovered.

Modelling movement growth

In order to understand the counterfactual impact of a marginal intervention to foster movement growth, it's helpful to first have a picture of what we expect to happen without that intervention. Generally this will be of the form: (i) the movement grows as more people become aware of it; (ii) as it grows more direct work is done; (iii) at some point the movement size plateaus.

I'm going to introduce a simple model of the space in which this is happening. This will allow us to analyse the effects of different kinds of intervention.

The Awareness/Inclination Model (AIM)

People will know a variety of different things about the movement, and have different attitudes towards it. As a modelling assumption we will compress these down to two dimensions: awareness (how much they know about the movement) and inclination (how favourably disposed they are towards the movement, or how favourably they would be if they knew more).

Then a person (at a fixed point in time) will be represented by a point in this space. Some examples are shown in Figure 1.

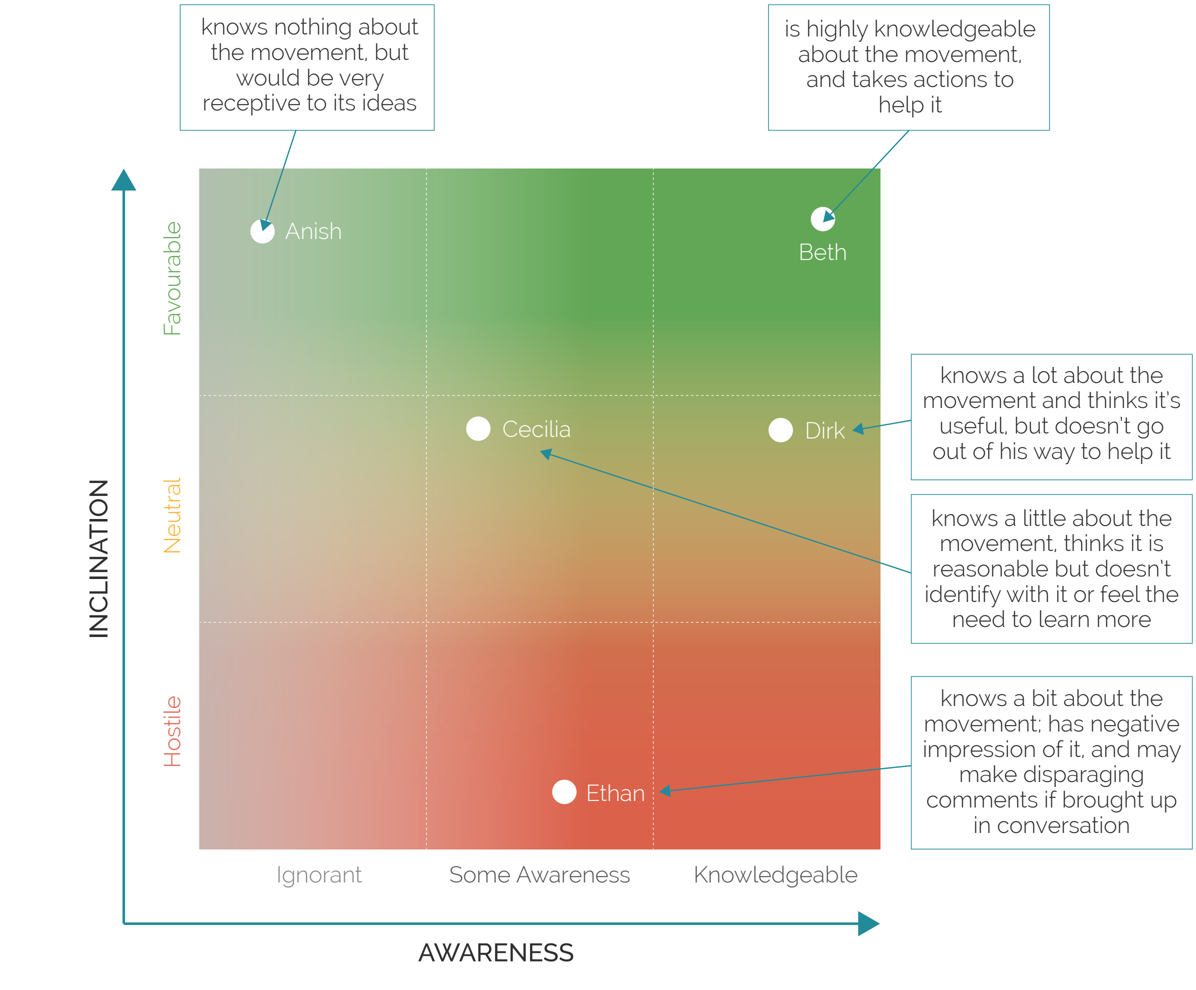

knows nothing about the movement, but would be very receptive to its ideas
is highly knowledgeable about the movement, and takes actions to help it
Anish
Beth
Cecilia
Dirk
Ethan
knows a lot about the movement and thinks it's useful, but doesn't go out of his way to help it
knows a little about the movement, thinks it is reasonable but doesn't identify with it or feel the need to learn more
knows a bit about the movement; has negative impression of it, and may make disparaging comments if brought up in conversation
INCLINATION
Favourable
Neutral
Hostile
Ignorant
Some Awareness
Knowledgeable
AWARENESS

Figure 1: Individuals in the AIM space

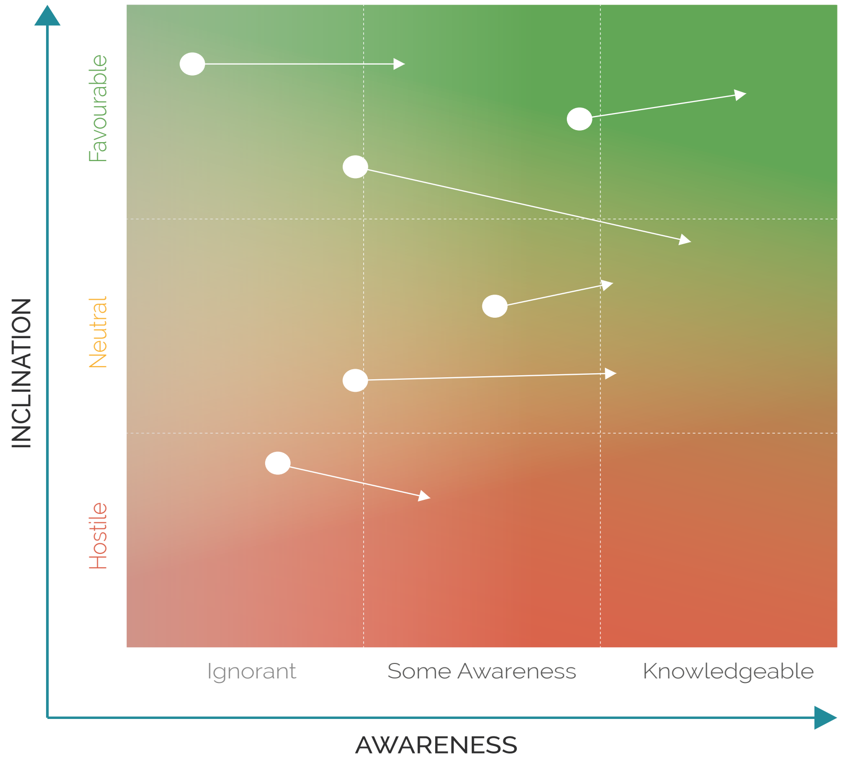


Figure 2: Example trajectories as individuals become more aware of the movement

Of course over time people may move around in this space. This will usually be driven by finding out more about the movement. So when movement occurs, we should expect it to be primarily in the left-to-right direction of the diagram. As they learn more, people may also become a bit more favourable or hostile to the movement. It seems to me that significant changes in inclination without significant increases in awareness are likely to be relatively rare during the growth phase. This is a substantive assumption which could be worth exploring, but the conclusions of the model do not require it.

Examples of some trajectories individuals could take through this space are shown in Figure 2.

Life-cycle of a movement

Early on in the life of an idea a large majority of the population will be ignorant of it. We cannot in general say how favourable they will be. Movements where the people on the left are clustered towards the top and away from the bottom have more potential for growth, and less potential to lead to a vocal opposition. It is the people in the top-right of our diagram who are both knowledgeable and favourably inclined who will actively pursue and promote the idea, and may be regarded as the movement built around that idea. This is illustrated in Figure 3.

As the movement starts to grow, some people will become more aware of it, moving along trajectories to the right and perhaps changing inclination slightly (Figure 4).

In our example, the growing awareness resulted in a substantially larger movement, as well as someone moving into the bottom-right corner of the diagram, becoming knowledgeable about the movement and also hostile to it. Why should this section matter? Consider what drives the increases in awareness

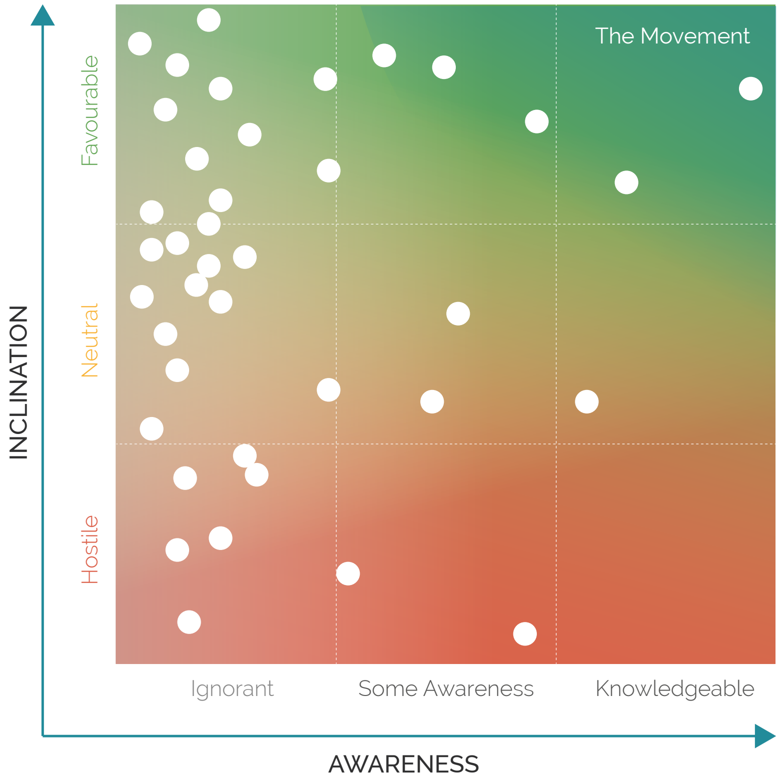


Figure 3: Early in the life of a movement

and changes in inclination. At the level of the model it seems that more people being knowledgeable and favourable to the movement will help to increase the spread of awareness, and that more of the knowledgeable people being favourable rather than hostile will help to nudge people's inclination in a positive direction. So having a larger, more active opposition to the movement may decrease the frequency

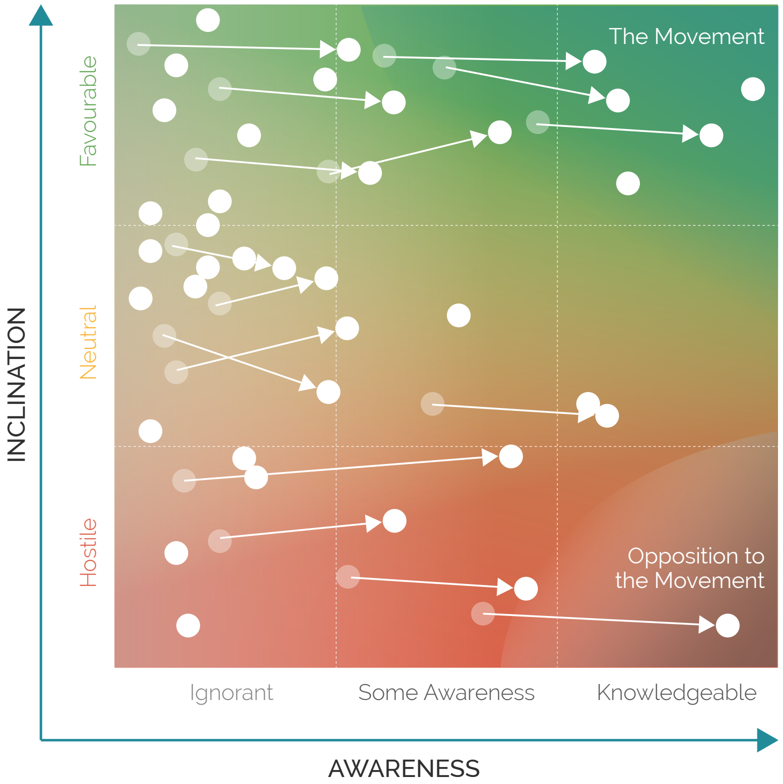
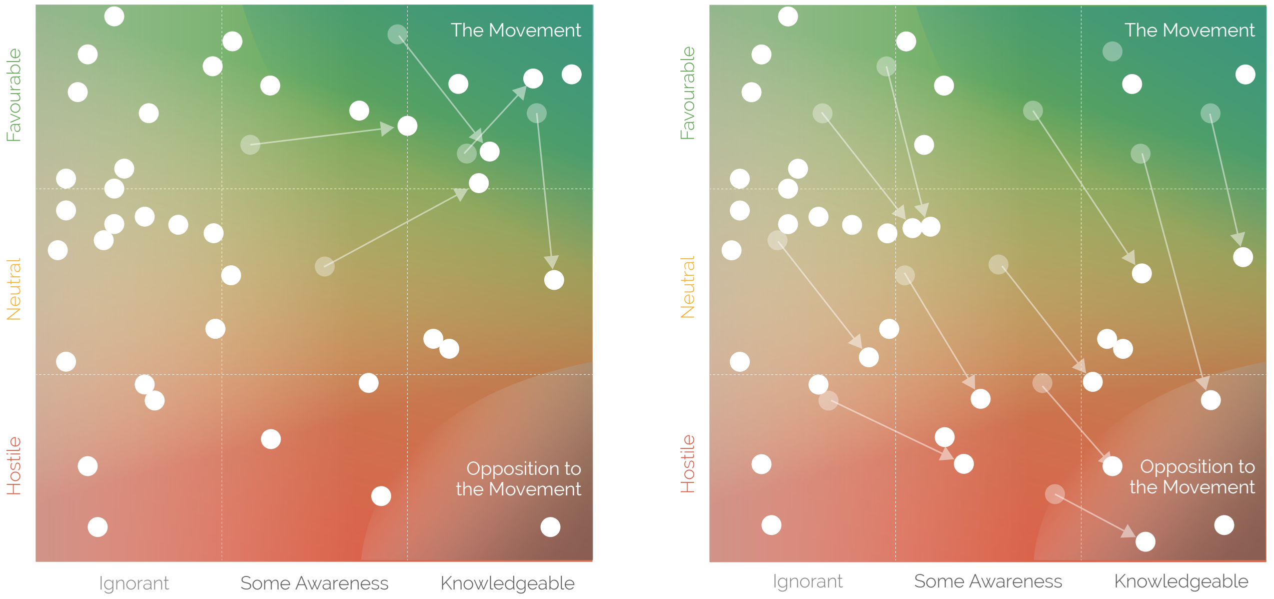


Figure 4: Growth of a movement

of positive inclination-shifts as people learn more (and increase frequency of negative inclination-shifts). It may also affect how the movement is able to influence public policy: a small highly favourable group with no opposition may be better able to shift policy than a larger group with a vocal opposition.

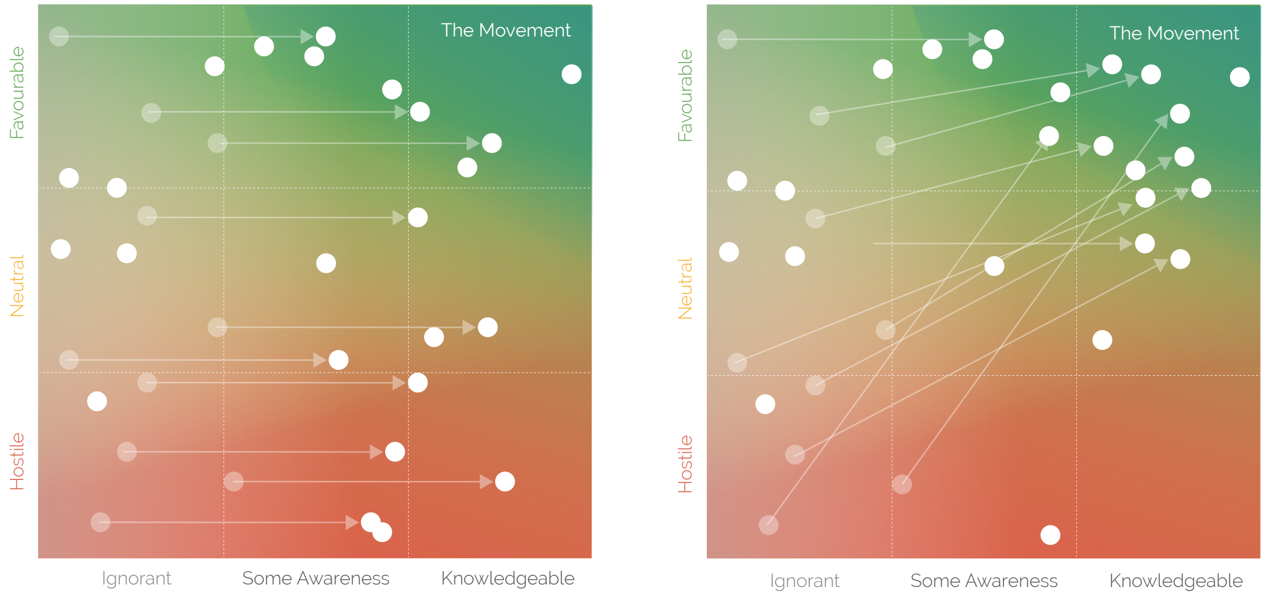
What is the long-term picture? There are four stable situations that can come about in the mid- to long-term as the number of people informed about the movement grows:



A - SATURATION

B - COLLAPSE

Figure 5: Scenarios where many people never become aware of the movement



C - CONTROVERSY

D - ACCEPTANCE

Figure 6: Scenarios where virtually everyone becomes aware of the movement

Scenario A – Saturation: There is a certain amount of work that should usefully be done on the issue at hand (perhaps each year, or perhaps in total). When the movement gets large enough, this work gets done, and there is no need to make further people aware; the movement is the correct size and more growth might even be bad. Existing people drift away from the movement as they see less need to be involved. e.g. fundraising for specific projects.

Scenario B – Collapse: Something happens to cause negative inclination towards the movement. This could be an inevitable consequence of increased awareness, or could be caused by some particular event. This affects both existing supporters and people who do not know about it. A subcategory of collapse is stagnation, where there is no real opposition but people lose their strong positive inclination, and nobody works to bring new people into the movement. In either case the negative inclinations stops the movement from growing, and it eventually collapses. e.g. Marxism

Scenario C – Controversy: In this scenario almost everyone knows a reasonable amount about the movement. A fraction of society is favourable enough to be actively involved, but many people are not in support. Most large movements can expect to be here. e.g. abortion

Scenario D – Acceptance: It becomes socially normal to support the movement. Almost everyone knows something about it and agrees.[72] e.g. ending the use of lead in consumer products

Effects of interventions

The marginal effect of an intervention may be quite different depending on which long-term scenario we are in. For this reason it's important to have a good understanding of which are the most likely. To help think about the effects of interventions, we'll consider the two intervention archetypes illustrated in Figure 7.

Intervention X – Publicity: This intervention focuses on raising awareness of the movement. It is consistent with the view that all publicity is good publicity. e.g. performing stunts that are widely reported.

Intervention Y – Advocacy: This reaches fewer people than Publicity, but is concerned chiefly with improving their inclination towards the movement. It is closer to brand management. e.g. writing a persuasive opinion piece for a newspaper.

Both of these interventions affect the population roughly uniformly; in practice many interventions may target subgroups.

Shifting timelines

The *Saturation* and *Acceptance* scenarios are perhaps the simplest to analyse. In both cases there is a

[72] The Acceptance scenario has a couple of variations: one where there is a positive feedback loop which increases inclinations towards the movement; a second where people's natural inclinations are positive enough that shifting right in the diagram is enough, and the movement is only limited by awareness. In the very long term Controversy could shift to Acceptance, for example as in the abolition of slavery.

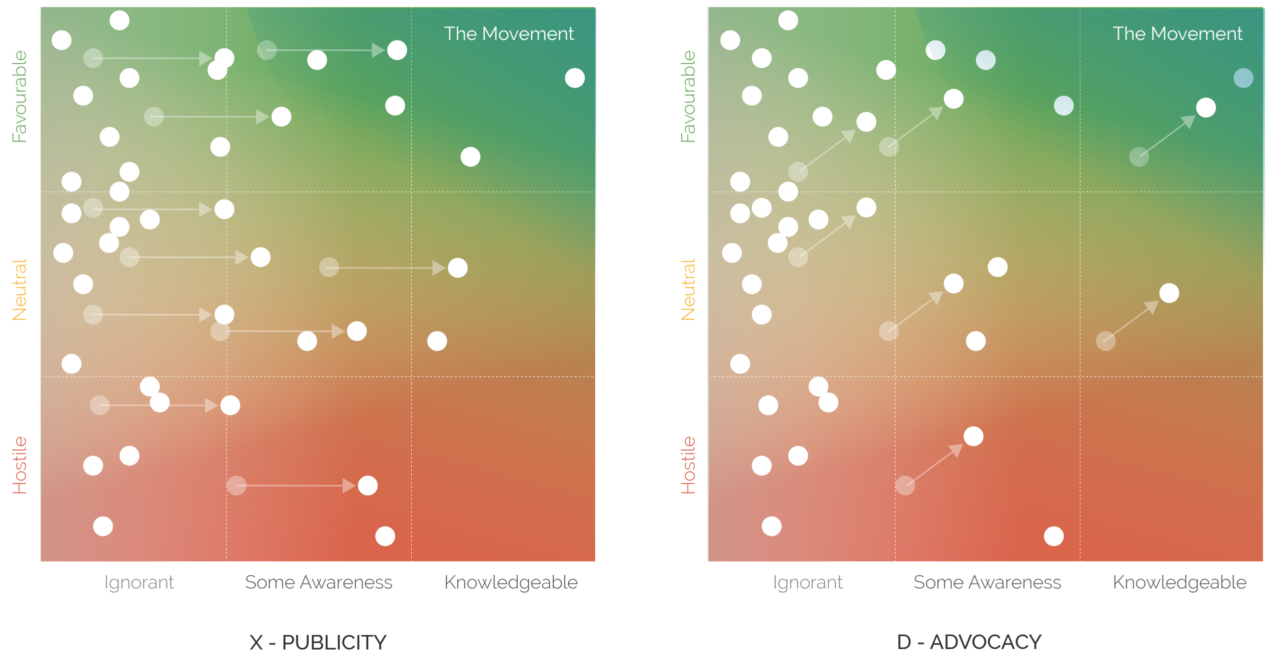


Figure 7: Two intervention archetypes.

natural eventual size of the movement, so marginal work on movement growth will not change that size, but change when it is realised. This suggests that the focus should be on increasing awareness rather than improving inclination: Publicity has a bigger effect than Advocacy on reaching the end state quickly. There is still a question of how good increasing awareness is compared to direct work.

The curve of movement size against time before intervention (black curve in Figure 8) thus has a horizontal asymptote. Suppose that we intervene to make the movement grow by k people. We will assume for now that our intervention just moves the entire curve forward in time (red intervention and blue curve in Figure 8).

What is the total effect of this? The growth of the movement is advanced in time by a certain amount a (see Figure 9). This will produce extra direct work on corresponding to the area between the blue and black curves.[73]

This area can be shown to be equal to a times the difference between current movement size and eventual movement size (see Figure 9 – so long as k is small the area of the extra grey triangle is negligible). If we value direct work equally regardless of when it happens, this gives a simple way of understanding the value of acceleration: it is only necessary to estimate the current rate of growth and the eventual movement size. Then:

Benefit = (eventual movement size – current movement size) * advancement period

For example, suppose the movement currently has 10,000 people in it, is growing at 10 people per day, and is projected to reach an eventual size of 100,000 people. Then an intervention which gains 10 extra

[73] Note that area in this diagram represents person-years in the movement; we are assuming that the direct work done is in proportion to this.

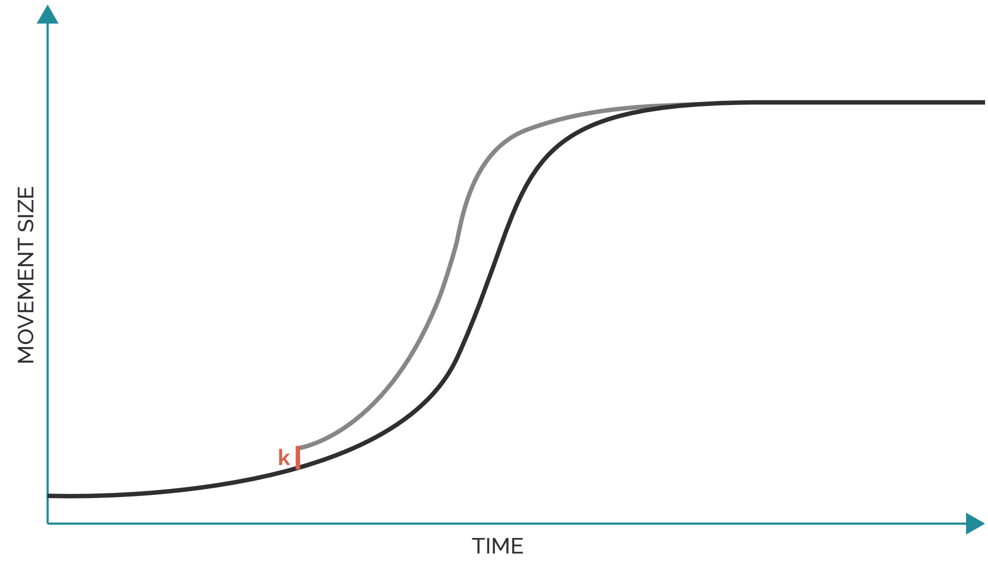


Figure 8: Intervention to accelerate movement growth

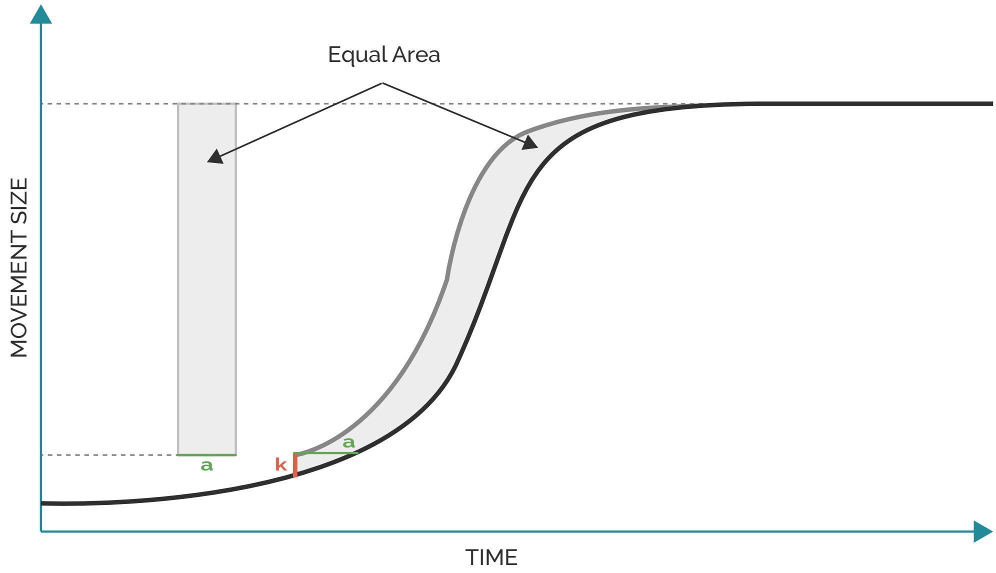


Figure 9: Assessing the impact of accelerating movement growth

new members beyond what would have happened anyway for the movement accelerates the process by a day, so long as it also advances the background awareness of the movement at the same rate.[74] This means the value of this advancement is the same as the direct work produced by 90,000 people in the movement for a day – which is the same as one person for 250 years. So the value of just one person joining can be estimated as having their future direct work for around 25 years.

Some caveats. A person being worth about 25 years seems reasonable, but this is a bit of a coincidence

[74] This corresponds to the assumption that we are simply moving the curve forwards. An example of an intervention which does not satisfy this assumption might be persuading 10 people who already think of themselves as almost members of Giving What We Can to fill in the membership form. However doing that together with sufficient outreach to refresh the pool of such people probably would satisfy the assumption. This means that outreach which only shifts people along at one part of the curve should be discounted by up to an order of magnitude compared to the direct measured shift.

of the numbers. The value is quite sensitive to eventual movement size. If the movement will only reach 20,000 people, each person added only equates to 2.5 years of direct work. If it will reach 10 million, each person added (while it's currently only 10,000 people) produces a whopping 2,500 years of direct work, because they pull forward the future hump of the distribution by a macroscopic amount. Uncertainty about eventual movement size will complicate this, but much of the expected value may be driven by scenarios where the movement gets very large, if these are realistic. In practice interventions won't advance all parts of the awareness process equally, so just counting new members is likely to overestimate the value (though probably by less than an order of magnitude). And if we value direct work differently at different times (or differently according to how much has been done), that would change the analysis, most likely in the direction of making direct work better relative to movement building.

For these scenarios it seems that movement growth can compare very favourably to direct work, and that Publicity is often better than Advocacy.[75] The major exceptions are when the movement has already done most of its growth, or if the movement is simply never expected to get large.

Shifting the long-term state

The Controversy scenario in some ways admits a similar analysis. Treating the eventual size of the movement as known and fixed, we can estimate how good accelerating movement growth is in terms of direct work. But an important difference is that we may be able to affect the eventual size of the movement. In Saturation and Acceptance, movement growth can be very important, and it is generally worthwhile to speed it up as much as possible by whatever means are most effective. In Controversy, long-term stability is achieved when everyone knows about the movement but people have different opinions about it. It may be possible to change this eventual distribution of opinions, and this could have a large effect.

For example suppose we have a movement of 10,000 people growing at 10 people per day, and projected to reach a size of one million people. Suppose there is an opportunity which would bring in 300 new people – accelerating growth by a month – but by being controversial, reduce the pool of people interested by 1%. At the current movement size, putting off 1% is just 100 people, so attracting an additional 300 seems like a great deal. But if the change persists and affects the stable level, this would be reduced by 10,000. Projecting twenty years past the plateau of movement size, the intervention will destroy more than twice as much direct work as it creates in this model.

Since we hypothesise that a major driver of improving inclination towards the movement is having lots of people with high inclination and few people who are opposed, in Controversy it may be more important to focus on improving inclination towards the movement as a first priority, and raising awareness only as a second priority. Advocacy is probably a preferable intervention to Publicity.

In Figure 10 we suppose we have a metric for inclination towards the movement such that the population

[75] This corresponds to the assumption that we are simply moving the curve forwards. An example of an intervention which does not satisfy this assumption might be persuading 10 people who already think of themselves as almost members of Giving What We Can to fill in the membership form. However doing that together with sufficient outreach to refresh the pool of such people probably would satisfy the assumption. This means that outreach which only shifts people along at one part of the curve should be discounted by up to an order of magnitude compared to the direct measured shift.

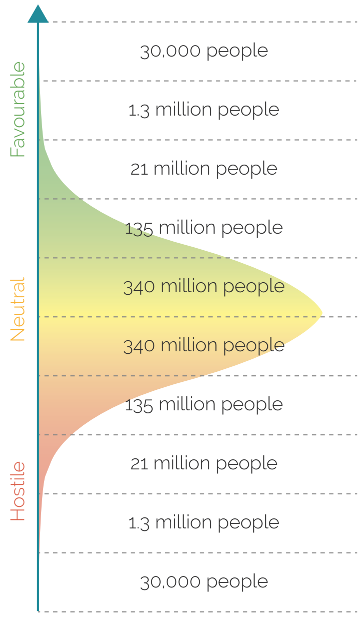


Figure 10: Distribution of level of inclination in a population of 1 billion people

is normally distributed.[76] The horizontal bands show standard deviations (z-scores), and the dots and numbers show how many people lie in each band, from a population of around a billion people.

Changing the threshold required for eventual participation in the movement can have a large effect size. For example, suppose in one scenario the movement only catches on in those who are 3.8 standard deviations towards favourable or higher, whereas in another the movement catches on in those who are 3.7 standard deviations towards favourable or higher. Shifting from the first to the second scenario seems like a small change in required attitude, which might be achievable by shaping the right positive view of the movement. But its effect is to add 35,000 people to the eventual size of the movement, on a base of 72,000 – almost a 50% increase! So if how positive people will eventually be towards the movement can be adjusted, that could be crucial (see Figure 11).

Interventions which change the likelihoods of different long-term scenarios may also be crucial. Changing scenarios is most plausible between Collapse and the other scenarios. This could have a very large total effect. Avoiding movement collapse may be difficult to do with positive interventions, but it becomes extra important to avoid events which could catalyse the creation of opponents of the movement and drive movement collapse.

Changing the movement

In the above, we've imagined that changing the framing around the movement can affect the size it grows to, but that the underlying message and the direct work associated are essentially the same. It's also possible that by choosing a slightly different movement to promote, for example by excising less popular ideas, we could change the growth.

This may be a powerful tool. Because the eventual movement size may be so sensitive to – among other factors – how appealing the ideas are (as in Figure 11), it could be valuable to look for opportunities to adjust and improve this. It's also a tool that should be used very carefully, however – the value of

[76] If necessary this might be a synthetic metric, in the same way that IQ is.

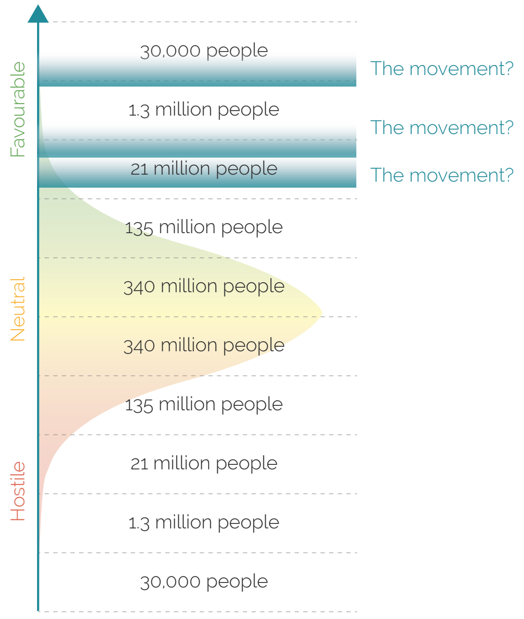


Figure 11: Possible eventual movement sizes – depending on how extremely favourable people have to be in order to become involved

spreading a different movement may be substantially different, and it could be too easy to throw the baby out with the bathwater by removing unpopular components.

Limitations and extensions

The model we have used necessarily makes a number of simplifications. I think these have generally been worthwhile, in allowing us to go further and faster through the analysis without distorting the picture too much. Nonetheless, it's worth being aware of possible limitations or shortcomings, and it may be that future work could explore relaxing these assumptions.

Compressed space of attitudes

Restricting to the 2-dimensional AIM space meant that we could not represent more complex attitudes, such as distinguishing between someone who has strong positive affect towards the movement but is not currently taking action on this, and someone who has much more mixed feelings but is currently pursuing some direct work. Much of the time I think these two dimensions suffice to capture the important distinctions, but there may be cases where they fail.

No loss of awareness

We assumed that people tend to move in AIM space only when presented with something external, and cannot move to the left. In fact people may drift around without interacting with others, and in particular may forget and so slowly lose awareness. It seems likely to me that the pictures presented are still roughly correct on average for growing movements, but further consideration of this point might show extra benefits to some interventions.

No different behaviour for distinct subgroups

Perhaps it is easier to affect inclination by large amounts in some groups such as students. The current model doesn't make any assumptions about this, but doesn't make it easy to model. An extension of the model could perhaps provide easier ways to think about heterogeneity in the population beyond the position in AIM space.

Growth is primarily endogenous

When we looked at the marginal effect of a larger movement today, we assumed that this would accelerate the movement growth process fairly uniformly. This is reasonable if growth of the movement is driven primarily by people in the movement. If growth is driven significantly by some exogenous factor (for example uptake of a new technology, reaction to political events, or flow from a related movement) then the long-term effect of growing the movement today could be much smaller. Changing the dynamics so that the long-term state alters could still be very important.

Applications

Types of intervention

Our example interventions Publicity and Advocacy were quite a long way removed from concrete actions. What might they look like in reality? Since the effects are all about changing minds (via awareness and inclination), interventions will tend to take the form of communication – although they could also include taking actions that can usefully be reported.

Example intervention might include:

- Writing an introductory article in a popular media venue. This would increase awareness across the population. Depending on style, it might also increase inclination, or be more controversial and sacrifice inclination for extra awareness.
- Writing a detailed articulation of an idea important of the movement, to publish on a website associated with the movement. This would increase awareness and inclination, primarily in a group near the top-centre of AIM space. The article you are currently reading is an example of an intervention which intends to be of this type.
- Direct work on the problem the movement is aimed at. This will not affect awareness, but improve inclination as people become more aware.

The last example may be surprising: we started out setting movement building against direct work, and are now using direct work as an example of movement building! What's going on here?

The answer is that while direct work has some movement growth effects, it is usually not a very effective way of movement building, so there is a meaningful choice between direct work and (effective) movement growth. However if the ratio of direct work to movement growth activities is too low, this could hinder the growth of the movement, either because it gives a natural line of criticism, or because it is harder to tell a clear story of what the movement is about. In such cases direct work may become the most effective form of movement growth at the margin – albeit with a slant towards direct work that can be communicated clearly or sets a good example.

As a final note, when choosing what to do, it's important to keep the concept of comparative advantage in mind. Even if you think a bit more movement growth is the most effective thing to be doing at the margin, if you have much better opportunities to pursue direct work, that could be better. Think about the portfolio of work done by those involved in the movement. Direct work should generally be done by those best at it, and movement building by those best at that. The importance of movement building relative to direct work sets the location of the split between these activities and shows what people with intermediate skill sets should be working at.

Effective altruism

The effective altruism movement is concerned with using evidence and reason to do the most good it can in the world. It is currently quite small (in the thousands of people, probably the low thousands) and growing rapidly (for example membership of Giving What We Can has roughly doubled each year).

Decision-relevant parameters

A key decision-relevant parameter is how big the movement will get. It seems that it would be hard to run out of effective things to do, so Saturation is impossible. Collapse, Controversy, and Acceptance are all possibilities. Movement collapse is probably unlikely, but it is very important to avoid.

Knowing how big the movement might get would tell us whether we're likely to end up in scenario Controversy or Acceptance, and if Controversy, what the scale of the eventual plateau might be. I think there is a lot of uncertainty around the eventual size. It could be valuable to do market research to estimate what the uptake rates could get to in different slices of society, although this is not straightforward.

If, as seems to me most likely, we are heading towards Controversy, it is also crucial how much we could hope to affect the eventual size of the movement. Again, this is hard to estimate, but might be done with market research, seeing how much effect different framings have on people's willingness to engage.

Implications

The higher we think the plateau is, the more we should care about movement growth relative to direct work. The more we think we can affect the level of the plateau, the more we should prefer quality to speed of growth, focusing more on inclination than awareness. And the higher the discount rate we place on direct work, the more we should care about direct work compared to movement growth, and the more about spreading awareness than improving inclinations.

My impressions are that it is possible the movement could get very large – millions of people within twenty to fifty years. Until we have ruled this out, it is important to keep movement growth a priority, and to weight improving inclinations more than increasing awareness in many trade-offs. Focusing on improving inclinations means that we should be careful with our framing, trying to explain how we help people to do things they already wanted to do.

Effective altruism has several facets, as people pursue direct work in multiple areas. Dealing with developing-world poverty has been perhaps the largest focus area of the movement. I think that at present movement growth is not significantly limited by a lack of direct work in poverty, and this work is not dramatically more useful now than in a few years. The implication of this is that it may be better at the margin if we put a few more resources into movement growth rather than direct work in poverty. Of course this statement has several caveats:

- This is by no means a claim that poverty isn't a hugely important issue, and one with great opportunities for direct work available. It may well be that the big success of effective altruism will be to help eliminate poverty. The claim is just that extra direct work today may have less total impact on poverty than extra movement growth today.
- This doesn't mean that we should collectively stop doing direct work in poverty altogether. This is important to continue: because it would be faithless to abandon the work; because it helps attract more people to such work; and because it helps us learn how to do good such work.
- Some individuals may be best increasing their direct work in poverty, because they have strong comparative advantage there or because they could be more effective advocates for effectively alleviating poverty if they could talk about their direct work in poverty.

Does the same argument apply to say that effective altruism should reduce its focus on direct work altogether to prefer movement building? Not entirely. I think that the movement's growth may be limited by credibility, by ability to assess cause areas or interventions, or by direct work in other areas such as engagement with policy. In general the direct work that is done will have an effect on the way that people perceive the movement, so may affect awareness and will certainly affect inclination.

Very approximate numerical estimates

Suppose the effective altruism movement currently has 3,000 people doing an average of $5,000 direct work per person per year (present movement size equivalent $15M/yr), and is growing by 10 people per day.

Great uncertainty over eventual movement size. Possibilities, excluding collapse:

- 10th percentile: $250 M/yr (50,000 people doing $5,000 per person)
- 50th percentile: $3B/yr (1M people doing $3,000 per person)
- 90th percentile: $50B / yr (50M people doing $1,000 per person)

Mean approx $10B/yr.

Adding a person to the movement ~ bringing this forward by 0.1 days. Adjust down to 0.025 days to account for the fact that there are lots of steps in getting people towards movement.

0.025 days * $10B/yr = $700,000 worth of direct work (excluding movement building effects from direct work).

Perhaps adjust this down a bit for future work holding less value than current work, and down a bit more for people doing otherwise useful work. Say $300k for getting a person into the movement.

That was for spreading awareness. What about improving inclination? (Numbers for this are if anything even more crude.)

Suppose increasing inclination of everyone who currently knows something substantial about the movement (say 10,000 people, which includes many people not in the movement) by 1σ leads to opinions being 0.1σ higher when movement reaches plateau, and that that change lasts 20 years.

Then causing a noticeable (0.2σ, corresponds to ~1.5cm height) shift in one person's opinion leads in expectation to 2×10^-6σ shift in population opinion. That leads to an expected ~200 people at 90th percentile, and around 25 people in expectation total, causing about $1.2M extra direct work. This should be adjusted down for later work being less valuable than early work – perhaps to $600k.

This is a large effect for improving opinions. It's probably most important for improving the opinions of people near the top and bottom of the inclination spectrum, as those are the ones who are likely to communicate most about the movement; therefore I suggest increasing this estimate for those people as well as for influential people, and decreasing it for those in the middle.

It might be better to model shifts in opinion as decaying with time and when spreading to new people. This will probably lead to a lower overall estimate of the value of changing opinions, since the decay will be biggest in scenarios where the movement gets very big and inclination is particularly important (a mock-up in Excel suggests it reduces impact by a factor of 4). Say $150k for a *noticeable shift in an opinion.*

All of these numbers are basically pulled out of my head and are neither carefully modelled nor terribly robust. Nevertheless, they are suggestive, and I'd be mildly surprised if they were wrong by much over an order of magnitude. It's important to remember that we excluded the movement growth effects of direct work in order to make the comparison, so direct work will tend to compare more favourably in practice than these suggest.

Key empirical questions for better numerical estimates:

- How big is the movement likely to get?
- How many people know about the movement now?
- How fast is the movement currently growing?
- How much do attitude shifts decay with time?
- How much are attitude shifts transferred across people?

Practical upshots

Here I'll try a little more interpretation, to discuss what this means when deciding between actions. These views are informed by the theory but not entailed by it:

- We should be very careful to avoid actions which could cause negative inclinations towards the movement for little gain. We should therefore avoid hostility where possible, and should try to avoid expressing opinions on controversial topics when we may be perceived as affiliated with the movement.
- When allocating resources to direct work, we should do so in such a way that we can demonstrate what we've done, and we can learn from the experience, even if there are associated overheads.
- We should communicate our successes clearly. If a significant amount of the value of direct work today comes from demonstrating credibility and persuading others, it's important to be able to draw

on this easily when relevant. (Note this is not the same as trumpeting our successes widely.)

- We should probably avoid moralising where possible, or doing anything else that might accidentally turn people off. The goal should be to present ourselves as something society obviously regards as good, so we should generally conform to social norms. Similarly presenting ourselves as on the side of the person addressed may help to avoid creating negative inclination shifts.
- We should try to reach good communicators and thought leaders early and get them onside. This increases the chance that when someone first hears about us, it is from a source which is positive, high-status, and eloquent.

Other areas and movements

How does this analysis translate over to other young movements, or other focus areas of effective altruism? For such linked movements, it generally seems that collapse and controversy are the most likely outcomes, that many people are relatively ignorant of it, and that the community of people who might care about it is much larger than the community that currently does. So for the same reasons as before it seems that movement growth is important, and that inclination may be more important than awareness. In cases where many of the eventual routes to impact are likely to involve policy, focusing on inclination and avoiding making opponents could be particularly important.

However, I think there is a significant disanalogy between effective work in global poverty and some more speculative areas such as existential risk reduction. Effective interventions in global poverty are well-demonstrated, so the growth of attention is not significantly limited by direct work. In contrast, the growth of attention for existential risk reduction is substantially limited by a lack of credible direct work in the area. Therefore, in order to foster movement growth, one of the highest value activities is to do direct work, and to set an agenda for direct work, that clearly helps with the problem. Choosing work that demonstrates credible ways to help, and explaining the value of this work, is probably more important than choosing the direct work with the biggest actual effect in the short term.[77] There may also be risks that need to be addressed in the short term, which could increase the discount rate on direct work, and so make direct work more valuable relative to movement building.[78]

[77] See Toby Ord

[78] See this FHI article

Further Resources

In recent years, the effective altruism community has created a lot of good content.

These articles are particularly useful for people who are already familiar with some of the core content, and want to understand the ideas in more depth.

Read

How can you figure out which global problem is most pressing?	80,000 Hours compare problems in terms of scale, neglectedness, solvability, and personal fit.
Hits-based Giving	Holden Karnofsky argues that high-risk projects might be the best bets.
Making Sense of Long-Term Indirect Effects	Indirect effects: very important, but hard to estimate.
Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity	Holden Karnofsky explains why the Open Philanthropy Project devotes unusually large resources to risks from advanced artifical intelligence.
Guide to working in AI policy and strategy	Miles Brundage gives a detailed account of how to best get involved in AI policy and strategy for 80,000 Hours.
Reducing Risks of Astronomical Suffering: A Neglected Priority	David Althaus and Lukas Gloor discuss risks of astronomical suffering and how to mitigate them.
Report on Consciousness and Moral Patienthood	Luke Muehlhauser looks at the literature on animal consciousness and moral patienthood.
Why Are US Corporate Cage-Free Campaigns Succeeding?	Lewis Bollard describes how a small group of campaigners convinced over 200 US companies to stop keeping chickens in battery cages.
Using Evidence to Inform Policy	How can evidence-driven research ensure that effective policies are implemented?
How Much Evidence Is Enough?	Argues that we should pursue interventions which have not been conclusively proved, but that we should go on collecting evidence beyond that point.
Reasons to Be Nice to Other Value Systems	Brian Tomasik argues that we should work together even when we value different things.
The value of coordination	To do good well, we need to work with others. But how can we coordinate effectively?
EA Community Building	What we might do to maximise the effective altruism community's impact.
Building an Effective Altruism Community	Why we might, and might not, want to help build the effective altruism community.
Effective Altruism Concepts	An encylopedia of interesting ideas in effective altruism.

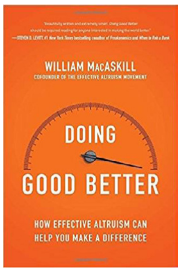
Books

80,000
HOURS
FIND A FULFILLING CAREER THAT DOES GOOD
BENJAMIN TODD
AND THE 80,000 HOURS TEAM

80,000 Hours

Benjamin Todd

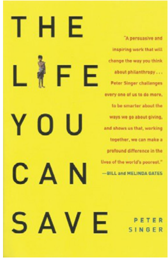
This handbook has focused on cause prioritization, but your career is the most important resource you have. We recommend 80,000 Hours' guide to finding an effective career.



Doing Good Better

William MacAskill

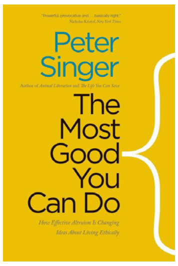
An introduction to some of the key principles of effective altruism.



The Life You can Save

Peter Singer

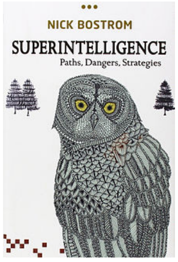
Singer argues that there is a moral imperative for the affluent to end the poverty in low-income countries.



The Most Good You Can Do

Peter Singer

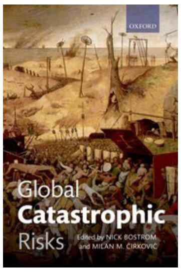
The Most Good You Can Do: This book explores the question of how to do the most good.



Superintelligence

Nick Bostrom

Bostrom sets out his view of how artificial intelligence is likely to develop, and the risks that it will pose.



Global Catastrophic Risks

Nick Bostrom, Milan M. Cirkovic

A survey of different risks that face life on earth.

Research websites

80,000 HOURS	80,000 Hours	Research into how to do the most good with your career.
ANIMAL CHARITY EVALUATORS	Animal Charity Evaluators	Working out how to most effectively improve animal welfare.
	Centre for the Study of Existential Risk	Studying how to mitigate risks of human extinction or civilizational collapse.
Charity Entrepreneurship	Charity Science Entrepreneurship	Assessing ideas for new high impact charities, and helping to bring them into existence
EA Concepts	EA Concepts	A rough conceptual map of the effective altruism research space.
Foundational Research INSTITUTE	Foundational Research Institute	Research into how to cooperatively and effectively reduce involuntary suffering.
Future of Humanity Institute UNIVERSITY OF OXFORD	Future of Humanity Institute	An academic institute focused on understanding how to improve the long-term future.
GiveWell	GiveWell	In-depth analysis of the best giving opportunities in global health.
CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE	Leverhulme Centre for the Future of Intelligence	Interdisciplinary research on how we can make the best fo the opportunities of artificial intelligence.
MIRI MACHINE INTELLIGENCE RESEARCH INSTITUTE	Machine Intelligence Research Institute	Foundational mathematical research to ensure that smarter-than-human artificial intelligence has a positive impact.
Open Philanthropy Project	Open Philanthropy Project	Lessons from an attempt to give away billions as effectively as possible.
SENTIENCE INSTITUTE	Sentience Institute	A think tank which explores the most effective strategies to expand humanity's moral circle.
Wild-Animal Suffering Research	Wild Animal Suffering Research	Research and advocacy aimed at identifying effective interventions to improve the wellbeing of wild animals.

Watch and Listen

In the 80,000 Hours Podcast, Rob Wiblin interviews experts from a range of fields about how best to tackle the world's most pressing problems.

Doing Good Better is a short podcast series which introduces effective altruism. It is centred around three questions: Why? How? and What?

Effective Altruism (6 minutes)

Beth Barnes on what we could achieve if we all gave effectively

Effective Altruism: A Better Way to Live an Ethical Life (1 hour 24 minutes)

William MacAskill debates Giles Fraser on Intelligence[2]

Take action

1 – Learn more

We know that this doesn't sound like a very practical first step. But it is difficult to work out how you can best contribute to solving the problems the world currently faces, and learning more is often the most useful place to start. For example, if you're interested in what you should do with your career, we recommend 80,000 Hours' Guide.

2 – Give to outstanding charities

One of the ways that you can most immediately help is to give to effective charities.

We recommend donating through our EA Funds platform. This allows you to set up one-time or recurring donations to a variety of philanthropic funds. Experts in that area will then guide your donation to the most effective organization they can find. You might also consider the recommendations of some specialized charity evaluators: for instance, GiveWell evaluates global poverty charities.

3 – Discuss and connect

Working out how you should help is worth puzzling over. Your career is probably the biggest resource you have for doing good. And since we all have different skills and experience, you will need to find a career that is a good personal fit.

While you can learn a lot online, discussing things with others is often the best way to figure out how you can best help. It also helps you find people who you could collaborate with on useful projects.

There are effective altruism groups around the world, which discuss these ideas, and run projects aimed at helping others effectively. You can find your nearest discussion group here. You can also see if there's an upcoming event near you.

You can also discuss ideas and projects online on the Effective Altruism Forum.

CENTRE FOR
EFFECTIVE
ALTRUISM

This document has been put together by the Centre for Effective Altruism. For more information on effective altruism, go to effectivealtruism.org