Transcript: Ezra Klein Interviews Holden Karnofsky

nytimes.com/2021/10/05/podcasts/transcript-ezra-klein-interviews-holden-karnofsky.html

The New York Times October 5, 2021

Holden Karnofsky, a co-founder of GiveWell, explores the ethical questions around charitable giving.

How to Do the Most Good

Holden Karnofsky, a co-founder of GiveWell, explores the ethical questions around charitable giving.

[MUSIC PLAYING]

ezra klein

I'm Ezra Klein, and this is "The Ezra Klein Show."

Over the past— I don't know — five, six years, I've been very influenced by the effective altruism movement. On one level, effective altruism is simple. It asks, how do we do the most good we can with the money and the resources we have? That turns out to be, one, a deceptively difficult question and, two, weirdly, one that we don't ask all that often, one that oftentimes you think people are asking and they are not.

But the difficult parts are maybe more interesting. How do you measure the most good? What about when you think something is good, but it cannot really be measured? Who defines good? Who verifies impact? How do you judge the value of, say, supporting art against the value of building housing for the poor?

Effective altruism has roots in the academy. Philosophers like Toby Ord and Will MacAskill and Peter Singer, they've been central in creating the movement. And importantly, they're central in the way the movement thinks and reasons. The culture of effective altruism, in my experience — and this is both its best and worst quality, in a way — can feel like a philosophy grad seminar that never ends.

By that, I mean it delights in taking the logic of its questions as far as it will go. It's unafraid, even ecstatic, to follow answers that strike others as very strange or unintuitive, sometimes even cruel. It's always, always questioning its own assumptions and everyone else's. It can, in my view, sometimes be performatively cold or logical in a way that's actually quite narrow about human flourishing. But as I've said, I have learned a lot from these thinkers.

What's interesting to me is that effective altruists tend to come out on two very different ends of what you might end up worrying about. Some take the analysis and end up worrying a lot about very provable, very, very well defined interventions, like distributing malarial bed nets, because those are the interventions we have studied with the best research. We've done randomized controlled trials. And we know they save lives cheaply.

And then others come out obsessed with much more speculative threats, like, say, killer artificial intelligence because if you start running the thought experiments, if you assign a moral weight to the future and all the potential human beings in the future, anything that could end humanity tomorrow really demands our attention today. If you can improve the likelihood of countless generations even a little bit, that really nets out to a big impact on the future.

No one represents the dual and sometimes warring impulses of effective altruism better, to me, than Holden Karnofsky. Holden was one of the co-founders of GiveWell, which measures the effectiveness of different charities and recommends the ones it is most confident can save lives cheaply. But then he spun out of that to found Open Philanthropy, which is more on the speculative side of effective altruism, the thought experiment side.

But because he's actually giving away money and making grants and shaping sectors and commissioning research, he and his team have to be pretty serious about how they approach questions that are not always considered seriously. They don't even always sound serious when they are spoken aloud. So there is a practicality, even a groundedness, to his analysis, even when it ends up in very, very strange places.

And more recently, Holden has been going to stranger places still. He launched a blog called Cold Takes. And on it, he's making the case, in piece after piece, that we live in maybe the most important century humanity will ever have, the most important we have had or will have, that the future could be wildly unlike the past. It is real mind expanding stuff.

And this episode — warning — will go to some very strange places. This is not our normal fare. But I'd urge you to follow it there. There's a concept Holden and I talk about here — worldview diversification, the practice of recognizing that we're not always sure how we should look at the present or the future. And we should keep a number of different possibilities, some of them even outlandish, alive in our minds. We should even keep them alive in our actions, rather than trying, as we so often do, to choose between them. This is something I try to do on the show. You can think of each episode as a little exercise in worldview diversification. And I try, as you know, to stay away from the choosing, except when we really have to. But it's particularly something we do in this episode. Is Holden right about some of the more wild predictions we discuss at the end? I don't know. On some level, I have a lot of trouble believing it, but I do think it's good to stretch the predictions you're willing to entertain, if only to help you see the present more clearly. As always, my email: ezrakleinshow@nytimes.com.

[MUSIC PLAYING]

Holden Karnofsky, welcome to the show.

holden karnofsky

Thanks for having me.

ezra klein

I've wanted to do this a long time, man.

holden karnofsky

Cool.

ezra klein

It's a minute in the making. All right, I want to begin at the beginning for you, before Cold Takes, before Open Phil, back in the GiveWell days. Tell me the story of why you started GiveWell.

holden karnofsky

Sure. So I was a few years out of college, and I wanted to give to charity. And I had the immediate thought that maybe I could find a website that would just tell me where to give to charity to sort of get the best deal. I think you could maybe think of what I wanted as a Wirecutter for charities, although there was no Wirecutter then, either. What I wanted was to kind of help as many people as I could with the money that I was giving. And I tried to find this, and I couldn't.

And my coworker, Elie Hassenfeld, and I were in the same boat on this. We were going through the same journey together. And we both decided to create the website that we wish existed, which is GiveWell, which is just public recommendations, telling people which charities can help their money help the most people possible.

ezra klein

So I looked for things like this in this period. And I found things that at least purported to be this. People might be familiar with Charity Navigator. There's a lot of cross-charity budget comparisons. Why weren't those what you were looking for?

holden karnofsky

Sure, so at the time we were doing this was around 2006, 2007. And the obsession in the charity evaluation world was around the overhead ratio, or how much of a charity's budget does it spend on the so-called programs versus the so-called overhead. This is an appealing thing in that you have to report these numbers to the I.R.S., so you can get this number for any charity. But it really doesn't seem all that relevant to me.

For example, if you have a charity and you decide to pay your top talent more or you get better information technology, and now you're doing way more good and helping way more people, is that overhead? Is that somehow a waste of money? Is that, somehow, you did the wrong thing? And so, we wanted something different. What we wanted was this question of, what charity is going to help the most people per dollar that I spend?

And a crucial difference is that we were not in a place where we had a charity we already liked, and we wanted to look it up and make sure it was legit or that it wasn't fraudulent. We wanted something different. We wanted recommendations of which charities out of all the many you could do the most good. And so kind of like how the Wirecutter will actually kind of tell you what five products to maybe buy, instead of you look up a product and they tell you if it's going to break.

ezra klein

On a note, the Wirecutter is now a New York Times company, and I've not paid Holden to make all these Wirecutter analogies. But getting at this, one of the reasons there was an interest in things like overhead spending is it is cross-comparable across charities. As you mentioned, that's a number reported to the I.R.S. There are a lot of charities. They do very different things. Some of them work on international poverty. Some of them work on public health. Some of them give money to ballet. How do you begin to narrow down and test which ones do the most good for the least money?

holden karnofsky

Yeah, so we started with some criteria. We wanted to find charities that were proven, cost-effective and scalable. Proven means that there's strong evidence that what they're doing helps people. Cost-effective means that they're helping a lot of people per dollar. And scalable means that there's room for more funding. It means that if you give more money, more people will be helped. So we're not just talking about a charity that did something great once and wants more money now.

Those criteria will not capture everything good. And I imagine we'll get to this, that I've later branched out into other kinds of giving that don't have necessarily all those criteria. But when you have those criteria, if you look at any charity and you ask whether it's really helping people, you've got an academic literature to review. Does education interventions, do those help children learn better? Do health interventions help people live longer? And what you can do is you can start looking for the programs that are the most proven and the most cost effective. And then you can find charities that do them.

So that's a way to narrow the field. And one of the things that we learned early in GiveWell, which is a lesson that's kind of stuck with me throughout my journey to where I am now, is that if you want to do the most good, in some ways, the worst place you can start is your neighborhood, your friends, your country, because if you look at the whole population, in this case, with GiveWell, there's just a lot of people living in countries that are extremely poor by U.S. standards, and your money can go a lot further there. You can help a lot more people with the same money if you're willing to broaden your moral circle and include more people in it.

ezra klein

Talk me through this tangibly. What is the first charity you find where there is really, really good evidence that it can help people for very little money? Where do you find the charity? And where do you find the evidence?

holden karnofsky

I could talk about the literal first we ever recommended, but instead, I'll talk about the first where it really started to come together, that we felt we were nailing these criteria. So that was the Against Malaria Foundation. And what they do is they basically distribute insecticide-treated bed nets to help reduce the burden of malaria. Now, insecticide-treated bed nets cost about $5. They often cover two people. Sometimes they cover one person. They last for several years. And they can kill mosquitoes and block mosquitoes so that people don't get malaria.

There's a large number of very rigorous, randomized controlled studies on bed nets. And if you look at the effect size, it looks like they are helping people very cost effectively in the sense of — we estimate that there's a death averted for every few thousand dollars that you can spend distributing bed nets, which is kind of incredible and has been very hard to beat that number at any point during GiveWell's journey.

So, that's the intervention: that's distributing bed nets. And then what we found is we've looked at several organizations that distribute insecticide-treated bed nets. And one of them was really tracking the whole process. They were providing documentation that the bed nets were actually handed out. They were providing shipping information of the bed nets arriving. And over time, they would add more data on going back and surveying people if they were still using the bed nets. So now you can put together this whole case. We have very strong evidence base. You have an amazing deal, in some sense, where a lot of people get help for each dollar you spend. It's very hard to do that much good helping people in the U.S. with donations. And then you have this organization that is really repeatedly carrying it out and that is providing all the information we need. And so it becomes this very exciting way to spend your money.

ezra klein

So I want to hold on this idea that giving to anti-malaria charities could save a life for a few thousand dollars, or specifically, according to the GiveWell website, $3,000 to $5,000. How does that compare to more conventional causes that people may be familiar with, like disaster relief or donating to a soup kitchen?

holden karnofsky

It's about the best we've been able to find, in some sense. We've looked high and low for ways to help people for small amounts of money. Some listeners might be a little confused when they hear $3,000 to $5,000 because they might have heard, well, I saw a charity that can save a life for $0.20. But at GiveWell, it's always been about rigorously investigating the numbers and subjecting them to all kinds of scrutiny and analysis. And when you really go hard on the analysis and you ask that the number be real, $0.20 to save a life is not something that I think you really have the opportunity to do with your giving. And $3,000 to $5,000, it's just empirically, it's been — I mean, I think that's just incredible compared to most things that are out there.

ezra klein

What are some of the other numbers you've come to when you've looked at other methods?

holden karnofsky

A lot of times, we don't put a number on it so much as we just put a bound on it, where we'll say, hey, the evidence here is not very strong. The intervention itself is very expensive. It might be like thousands of dollars just to put a person in a program that might or might not be helping them at all.

And so, a lot of times, it's more a matter of saying, this doesn't have a strong enough case, or it's just intuitively too expensive. It's not going to match that other figure.

ezra klein

So one of the critiques of GiveWell, the GiveWell model, is that, to use the old line about economics, it only looks under the light, right? It looks for its keys, but only where the lamppost is. Because there are a lot of things that could be good in the world, and they don't have a very — it is hard to run a randomized controlled trial of them, because they only happen once, or they are effecting something more diffuse.

So what was the thinking behind demanding this very high level of empirical proof, which, on the one hand, can say, here's where my dollar is going, but on the other hand, you might say, yeah, but the leverage on that dollar is smaller because I'm only trying to save the one life as opposed to influence something, say, 10,000 or a million by averting a war or changing civil service or whatever it might be?

holden karnofsky

Right, right. I have a bunch of thoughts on this. The original thinking was just very pragmatic. It's just a start-up saying, let's do what we can do. I think sometimes, you do want to start under the light. If there's a big area under the light, well, if your keys happen to be there, you're going to find them a lot more quickly. Maybe you should start there. So I think in some ways, it's not always such a bad thing to do.

Another intuition we had is just that a lot of things that people try to do to help people are just very speculative. They're often based on having certain feelings about the world or feeling like certain things kind of just feel supportive to do. And we thought that things that are really supported by evidence, where you can really drill down and see how much money's going there, might actually be systematically better because they are kind of — they're optimized in a different way.

Or another way you could think of this is the less you know about some intervention, the more you might expect that it's going to just be the average thing you can do. And when you can create a strong evidential case that your money is doing incredible things, that might actually be better than other options.

The final thing I'll say, though, is that I, at least, have branched out a lot since then, so I now run a different organization or co-C.E.O. run a different organization called Open Philanthropy that does not have the same requirement that everything be totally proven based on evidence. And an interesting thing is that we've been looking for things that are better than GiveWell's top charities. And it's been really hard. It's really been surprisingly hard, yeah.

ezra klein

I'm going to hold you there because we're moving to Open Phil. But I want to do this a little bit more slowly.

holden karnofsky

Sure.

ezra klein

You're co-running GiveWell with Elie. It's going well. GiveWell became a big player on the block very quickly. I've given a lot of my money through GiveWell over the past five, 10 years. What makes you decide to split off and start a new organization?

holden karnofsky

Well, we met Cari Tuna and Dustin Moskovitz. Dustin is one of the co-founders of Facebook and also Asana. And they were trying to give away their fortune in a way that would help the most people possible. And we just felt that a different approach might be called for. When you have a public website making recommendations to anyone and everyone versus working with one family that's giving away a huge amount of money, the second one kind of starts to put new options on the table. And so, Open Philanthropy still recommends a lot of donations to give those type of charities. And Cari and Dustin still give a lot of money there. But there have been other things that may be higher risk or may be more in the mode of what I call hits-based giving, which is that if you can get the occasional success that is a really huge win, that might make up for a lot of donations that don't quite have the effects they wanted to. And so, that's what caused us to kind of pivot. And we started Open Philanthropy as a project within GiveWell, and it eventually spun out.

ezra klein

Before you begin Open Philanthropy, you undertake this big study of the history of the philanthropy space.

holden karnofsky

That's right.

ezra klein

Tell me a bit about that study and what you learned from it.

holden karnofsky

When we started working with Cari and Dustin, I wanted to understand a little bit more about how big philanthropies in the past, what they had accomplished. And I didn't know what I'd find. I wouldn't have been surprised if I learned that, actually, they'd never accomplished anything, and we should go back to doing the simplest stuff we can, because it's that hard to help people.

But that's not what I learned. What I learned is that there have been incredibly impressive successes from philanthropy that I think rank up there with the most important events for human welfare in the last 100 years or so. We actually at Open Philanthropy, we've now named about half our conference rooms — each one is named after a philanthropic success story.

ezra klein

So what are they named?

holden karnofsky

One of my favorites is called Green Revolution. And that was when the Rockefeller Foundation funded Norman Borlaug and others to research improving crop yields in Mexico. If I'd been around at the time, I doubt I would have said, oh, improving crop yields in Mexico, that's going to be the biggest success that's ever had by philanthropy.

But it may have been because this is generally now credited with kicking off the Green Revolution, where a bunch of countries went from importing to exporting food. And it's credited with saving a billion people from starvation. Norman Borlaug ended up winning the Nobel Peace Prize because it turns out that this was a very scalable improvement in agricultural productivity. So once they had these crops, they could just breed them anywhere. Agricultural productivity took off in a lot of poor countries. And that kickstarted all this economic growth.

Another one of my favorites that I'll just throw in is the pill, the common oral contraceptive for birth control. The work was funded by a feminist philanthropist named Katharine McCormick. And at the time, this is the kind of thing that wasn't going to get funded by the government because it was

controversial. And in fact, they weren't able to advertise the pill as birth control originally. Instead, the warning label was the advertisement. They had to put on a warning label that it could prevent pregnancy.

So it was an example of philanthropy being ahead of the curve, doing something that was controversial. But they ended up with something that was transformative and was a huge moment for feminism and human welfare because they were willing to be a little controversial like that.

ezra klein

One thing I noticed about both of the case studies you used there as examples of huge wins is they're technologies. The Green Revolution are new kinds of crops. The pill is a medication. I don't think most philanthropy is about seeding and staking new technological development. I'm not saying none is — I, in fact, I have a friend who's working on that kind of thing right now — but not most of it. So is that a problem? Is that one of the places where Open Phil begins to diverge a view that you should actually be doing product development through philanthropy?

holden karnofsky

Technology is not something I would say is neglected by philanthropy. You're probably right that most philanthropy is not technology philanthropy. There is a lot of philanthropy in science, especially today. It's become a very fashionable thing to put money into.

But I would say it does reflect this idea that if you want to do really incredible things that have massive scale and help tons of people, having innovation, having new technologies that can be copied and used freely by anyone, is a great way to get leverage. It's a great way to just have really big effects. And if you don't have any theory in your philanthropy of how you're getting massive leverage and affecting huge numbers of people, then you may be better off with the most straightforward, cost-effective, proven stuff.

ezra klein

Well, let's talk about how to put that kind of theory into practice. Give me some examples of what you end up funding through hits-based giving.

holden karnofsky

I mean, one example is just we're the largest funder in the world of farm animal welfare. That's both attempts to develop alternative foods that can reduce meat consumption, but also corporate campaigns to try and put pressure for better treatment of animals. Over the last few years, basically, every major grocer and fast food company in the U.S. has pledged to go cage free. We've been funding a lot of the work that led to that that was already going when we came, but we've tried to speed it up. And we've also taken that work global and been funding a lot of corporate campaigns globally.

And so, while we are a fairly large philanthropy, we're not the largest, but we're the largest funder of that work because that's one of these things where I think we might look back hundreds of years from now and say, that was the greatest moral issue of our time. That was this unacceptable treatment of these creatures we've now decided are kindred creatures that we should care about. But at this time, this is just not an issue that really tends to fire many people up. And so, we're doing something others won't do. And I believe there's been a lot of impact and there's been a lot of difference made because we're willing to go into kind of a weird cause.

Some other examples, I think we were the first major institutional funder of the YIMBY movement. This is the attempt to advocate for less restrictions on building housing to make housing more affordable. And this is, again, just something that was kind of new and weird and has now become a nationwide movement. And that was not how it was when we started funding it.

We funded macroeconomic stabilization policy. So this is a bit of a wonky one, but how does the Federal Reserve prioritize full employment versus controlling inflation? We believe that this is one of the most important things in the whole world for the welfare and bargaining power of the working class. And yet, it's an issue that people often ignore. They think of it as a technocratic issue. Whatever, the experts at the Federal Reserve will decide what to do. They don't see it as an area for philanthropy. And I think we are one of the only philanthropists in that area when we came in and still now. But I think it's tremendously important and we funded a lot of analysis and even advocacy on how to trade off full employment and controlling inflation.

A final example is we've had a biosecurity and pandemic preparedness program since about 2015. And I'm certainly not going to say I've been happy with the preparation response for COVID. But I think it could have been worse. And I think the organizations that have played important roles, by the time we all knew about COVID, it was too late to come in and support them for years and help them be in a solid position and build up a deep bench in a lot of expertise. And it was back when that was kind of an unusual cause for philanthropy to be in that we were supporting all that work.

ezra klein

Let me ask about a couple of these, though. Macroeconomic stabilization is an interesting one because one way of asking that question is, why do you think you all understand macroeconomic stabilization better than the Fed and others? You're a young organization. You're a young guy. You're not an economist, nor most of the people who work for you. A lot of people worked on this for a long time. You're coming in and saying there's a huge untapped opportunity here. I would understand for a very explicitly political organization to come in and say, oh, we think the Fed has gotten it wrong. We want more full employment. But what is the difference you have convinced yourself you can make on it?

holden karnofsky

Well, it's important to understand the general structure of Open Philanthropy is that we consider our expertise in finding causes that are important, neglected and tractable. Tractable means there's something for us to do. And so we try and find the right problems to work on. That's what we consider our comparative advantage. And then when we are doing work, we are hiring and we are funding experts.

And so, this is not about Holden going and learning all about macroeconomic policy and then going and explaining to the Federal Reserve that they've got it wrong. That's not what happened. We funded groups that have their own expertise, that are part of the debate going on. There are experts on both sides. But we funded a particular set of values that says full employment is very important if you kind of value all people equally and you care a lot about how the working class is doing and what their bargaining power is.

And historically, the Federal Reserve has often had a bit of an obsession with controlling inflation that may be very related to their professional incentives. And so we do have a point of view on when there's a debate among experts, which ones are taking the position they're taking, because that's what you would do if you were valuing everyone and trying to help everyone the most, versus which you're taking position for some other reason. So we didn't roll our own macroeconomic policy insights. We funded experts, we funded think tanks. But we do have a point of view on what kind of values should be driving that expertise.

ezra klein

I think something striking about that list is the sheer diversity of things you all fund. Not only in terms of causes but categories of causes. And this gets to what I think of as one of the most interesting things Open Philanthropy does, which is the way you intentionally divide up your giving portfolio into

buckets based on really different ethical, arguably even metaphysical, assumptions. So tell me about worldview diversification.

holden karnofsky

I need to start with the broader debate that worldview diversification is a part of. At Open Philanthropy, we like to consider very hardcore theoretical arguments, try to pull the insight from them, and then do our compromising after that. And so, there is a case to be made that if you're trying to do something to help people and you're choosing between different things you might spend money on to help people, you need to be able to give a consistent conversion ratio between any two things.

So let's say you might spend money distributing bed nets to fight malaria. You might spend money getting children treated for intestinal parasites. And you might think that the bed nets are twice as valuable as the dewormings. Or you might think they're five times as valuable or half as valuable or 1/5 or 100 times as valuable or 1/100. But there has to be some consistent number for valuing the two.

And there is an argument that if you're not doing it that way, it's kind of a tell that you're being a feel-good donor, that you're making yourself feel good by doing a little bit of everything, instead of focusing your giving on others, on being other-centered, focusing on the impact of your actions on others, which you can get from there to an argument that you should have these consistent ratios.

So with that backdrop in mind, we're sitting here trying to spend money to do as much good as possible. And someone will come to us with an argument that says, hey, there are so many animals being horribly mistreated on factory farms and you can help them so cheaply that even if you value animals at 1 percent as valuable as humans to help, that implies you should put all your money into helping animals.

On the other hand, if you value them less than that, let's say you value them a millionth as much, you should put none of your money into helping animals and just completely ignore what's going on factory farms, even though a small amount of your budget could be transformative.

So that's a weird state to be in. And then, there's an argument that goes, but even more than that — and this idea is called long-termism — if you can do things that can help all of the future generations, for example, by reducing the odds that humanity goes extinct. Then you're hoping even more people. And that could be some ridiculous comic number that a trillion, trillion, trillion, trillion, trillion lives or something like that. And it leaves you in this really weird conundrum, where you're kind of choosing between being all in on one thing and all in on another thing.

And Open Philanthropy just doesn't want to be the kind of organization that does that, that lands there. And so we divide our giving into different buckets. And each bucket will kind of take a different worldview or will act on a different ethical framework. So there is bucket of money that is kind of deliberately acting as though it takes the farm animal point really seriously, as though it believes what a lot of animal advocates believe, which is that we'll look back someday and say, this was a huge moral error. We should have cared much more about animals than we do. Suffering is suffering. And this whole way we treat this enormous amount of animals on factory farms is an enormously bigger deal than anyone today is acting like it is. And then there'll be another bucket of money that says, animals? That's not what we're doing. We're trying to help humans.

And so you have these two buckets of money that have different philosophies and are following it down different paths. And that just stops us from being the kind of organization that has stuck with one framework, stuck with one kind of activity.

ezra klein

Before we move on, I want to unpack this a little bit more. So let's focus in on animals for a minute. You alluded to the fact that even if you assign a very low moral worth to animals or to their suffering, 1 percent or 0.1 percent of that of a human, that it ends up adding up to quite a lot. Can you run through that math for me and its implications?

holden karnofsky

Well, the math would be that — I mentioned before that if you're distributing insecticide treated bed nets, you might avert the death of someone from malaria for a few thousand dollars, which is pretty amazing. And it's going to be very hard to find better than that when you're funding charities that help humans. However, with the farm animal work, for example, the cage free pledges, we kind of estimated that you're getting several chickens out of a cage for their entire lives for every $1 that you spend.

And so this is not an exact equivalence, but if you start to try to put numbers side by side, you do get to this point where you say, yeah, if you value a chicken 1 percent as much as a human, you really are doing a lot more good by funding these corporate campaigns than even by funding the bed nets. And that's better than most things you can do to help humans. Well, then, the question is, OK, but do I value chickens 1 percent as much as humans? 0.1 percent? 0.01 percent? How do you know that? And one answer is we don't. We have absolutely no idea. The entire question of what is it that we're going to think 100,000 years from now about how we should have been treating chickens in this time, that's just a hard thing to know. I sometimes call this the problem of applied ethics, where I'm sitting here, trying to decide how to spend money or how to spend scarce resources. And if I follow the moral norms of my time, based on history, it looks like a really good chance that future people will look back on me as a moral monster.

But one way of thinking, just to come back to the chickens question, one way of thinking about it is just to say, well, if we have no idea, maybe there's a decent chance that we'll actually decide we had this all wrong, and we should care about chickens just as much as humans. Or maybe we should care about them more because humans have more psychological defense mechanisms for dealing with pain. We may have slower internal clocks. A minute to us might feel like several minutes to a chicken.

So if you have no idea where things are going, then you may want to account for that uncertainty, and you may want to hedge your bets and say, if we have a chance to help absurd numbers of chickens, maybe we will look back and say, actually, that was an incredibly important thing to be doing.

ezra klein

I want to note something here because I think it's both an important point substantively but also in what you do. So I'm vegan. Except for some lab-grown chicken meat, I've not eaten chicken in 10, 15 years now — quite a long time. And yet, even I sit here, when you're saying, should we value a chicken 1 percent as much as a human, I'm like, ooh, I don't like that.

To your point about what our ethical frameworks of the time do and that possibly an open-field comparative advantage is being willing to consider things that we are taught even to feel a little bit repulsive considering, how do you think about those moments? How do you think about the backlash that can come? How do you think about when maybe the mores of a time have something to tell you within them, that maybe you shouldn't be worrying about chicken when there are this many people starving across the world? How do you think about that set of questions?

holden karnofsky

I think it's a tough balancing act because on one hand, I believe there are approaches to ethics that do have a decent chance of getting you a more principled answer that's more likely to hold up a long time from now. But at the same time, I agree with you that even though following the norms of your time is certainly not a safe thing to do and has led to a lot of horrible things in the past, I'm definitely nervous to do things that are too out of line with what the rest of the world is doing and thinking.

And so we compromise. And that comes back to the idea of worldview diversification. So I think if Open Philanthropy were to declare, here's the value on chickens versus humans, and therefore, all the money is going to farm animal welfare, I would not like that. That would make me uncomfortable. And we haven't done that. And on the other hand, let's say you can spend 10 percent of your budget and be the largest funder of farm animal welfare in the world and be completely transformative. And in that world where we look back, that potential hypothetical future world where we look back and said, gosh, we had this all wrong — we should have really cared about chickens — you were the biggest funder, are you going to leave that opportunity on the table? And that's where worldview diversification comes in, where it says, we should take opportunities to do enormous amounts of good, according to a plausible ethical framework. And that's not the same thing as being a fanatic and saying, I figured it all out. I've done the math. I know what's up. Because that's not something I think.

ezra klein

I'm struck by that. I really like worldview diversification as a way of thinking about things. And I think it's also relevant as an individual practice. Something I see in my travels around the world, the internet, is people are very intent. Even if they would not say they are 100 percent confident in their worldview, their political ideology, their whatever, they are really interested in making it dominant against all comers. So, just tell me a bit about organizationally, intellectually, the discipline of maintaining a certain level of agnosticism between worldviews whose differences you can't really answer.

holden karnofsky

So one of my obsessions is applied epistemology, which is like just having good systems for figuring out what your beliefs are in kind of an overwhelming flow of information that is today's world. And I think one of the tools that some people use for it that I find really powerful and I'm going to write about is what I call the Bayesian mind-set, which is this idea that when you're uncertain about something, you can always portray your uncertainty as a number. And you can portray it as a probability. There's thought experiments. There's tools for doing this. You can say, instead of something is true or false, that it's 30 percent. And you can look back later and you can see if things that you said were 30 percent likely come through 30 percent of the time. I think this is a very powerful framework. And using it can often get you out of the headspace of believing that things are true or false and just having degrees of belief in everything and often taking something very seriously, even when you think it probably won't happen, just because it's important enough and it has a high enough probability that it deserves your attention.

And on the other hand, I think this framework sometimes can take people back into a state of fanaticism, where you might say, hey, here's something that would be a really huge deal. And it's at least 1 percent likely. So that means it should be the only thing I think about. It should be my obsession. It's like the examples I was giving before. And that, I think, just lands you in a similarly dogmatic place.

And so, Open Philanthropy is kind of operating two levels of uncertainty. It's often using this Bayesian mindset. But when the Bayesian mindset brings you to this implication that you'll have to be all in on one thing or another, we'll say no to that, too. And then we'll just go to another level of diversification. And we'll have different buckets with different philosophies on the world.

ezra klein

I want to pick up on the fanaticism component. And I'm not accusing anybody here of fanaticism. But one of my critiques of the effective altruist world is that it can get very obsessed by that conversion number you were talking about a minute ago. And in particular, I think, it's a culture as it has matured a bit more. There's now an aesthetic, sometimes, of being willing to take the most hard-hearted logic experiment seriously and show that you're the real effective altruist because even though it sounds like a kind of terrible thing to do, you ran the math, and it's not.

And the way I'll put this is, Will MacAskill, who's a philosopher and was a founder of the effective altruist movement, used to have this thought experiment where there's a building on fire. And there's a family in one room who could die. And then there's another room — or I think it was, actually, an attached garage or something — that has a bunch of very expensive art in it. What do you save? And the point of the experiment originally was you should, of course, save the family. And he was making the meta point that many people are donating to museums, instead of to malarial bed nets. I think now, a lot of effective altruists would answer it the other way, because the point is, well, if that art is worth $500,000 and you can turn that $500,000 into x number of malarial bed nets, that saves more than five lives. And so, of course, you need to do that. And I think that gets you into pretty dangerous territory. But I'm curious how you think about those questions.

holden karnofsky

I do agree that there can be this vibe coming out of when you read stuff in the effective altruist circles that kind of feels like it's doing this. It kind of feels like it's trying to be as weird as possible. It's being completely hardcore, uncompromising, wanting to use one consistent ethical framework wherever the heck it takes you. That's not really something I believe in. It's not something that Open Philanthropy or most of the people that I interact with as effective altruists tend to believe in.

And so, what I believe in doing and what I like to do is to really deeply understand theoretical frameworks that can offer insight, that can open my mind, that I think give me the best shot I'm ever going to have at being ahead of the curve on ethics, at being someone whose decisions look good in hindsight instead of just following the norms of my time, which might look horrible and monstrous in hindsight. But I have limits to everything. Most of the people I know have limits to everything, and I do think that is how effective altruists usually behave in practice and certainly how I think they should.

ezra klein

What do you think the limit of that actual thought experiment is, of the just convert lives into money? You can save x number of lives for x number of money. And so if you get more money by getting the money as opposed to saving the lives, you should do it.

holden karnofsky

I think there's a lot of problems with that argument. And I could sort of go into them. So there's things about setting norms. There's things about following rules so that you don't want to be the kind of person who is constantly behaving in strange, unexpected ways and screwing over people around you because you've got this strange mathematical framework that's going on. So I think there's a bunch of things that are wrong with running in and saving the painting.

But I think I also just want to endorse the meta principle of just saying, it's OK to have a limit. It's OK to stop. It's a reflective equilibrium game. So what I try to do is I try to entertain these rigorous philosophical frameworks. And sometimes it leads to me really changing my mind about something by really reflecting on, hey, if I did have to have a number on caring about animals versus caring about humans, what would it be?

And just thinking about that, I've just kind of come around to thinking, I don't know what the number is, but I know that the way animals are treated on factory farms is just inexcusable. And it's just brought my attention to that. So I land on a lot of things that I end up being glad I thought about. And I think it helps widen my thinking, open my mind, make me more able to have unconventional thoughts. But it's also OK to just draw a line. I think it's OK to look at this art thing and say, that's too much. I'm not convinced. I'm not going there. And that's something I do every day.

[MUSIC PLAYING]

ezra klein

We've been talking a lot here about how to value animals, but the other big worldview here is long-termism, which has to do with valuing future human lives. So tell me more about that worldview.

holden karnofsky

So the basic idea of that worldview is that if you can do something today that affects all of the future generations, then you have helped a truly mind numbing number of people. We don't know how many people. We don't know how many people will live in the future. But it could be an extremely large number.

Most things that we can do today are not the kind of thing that we have any reason to believe will still matter a billion years from now. But some of them could be. Climate change could be. Climate change is an example of something that could really, in theory, could imaginatively knock humanity off course forever. And causing it to be less likely that this happens could be the kind of thing that matters for every future generation.

And so, long-termism says there are all these people who don't have any voice today in the actions we're taking that affect them. And so, why don't we take the actions that will still matter a billion years from now? Because they'll have affected that many people, and maybe that's the way to do the most good.

ezra klein

So one of the critiques of long-termism is it quickly gets you into this kind of mathematical moral blackmail, where, well, if you say, because in the future, human beings could spread throughout the galaxies and there could be a trillion of us, over and over again, there could be a trillion of us, so if you give the future human lives 1 percent of the weight of a current human life or 0.1 percent, sort of anything that makes that future more likely to happen is just an astonishingly good investment that outweighs anything you can do for people today. How do you think about that?

holden karnofsky

I have a few ways of thinking about this. One way comes back to worldview diversification again. So what we aren't trying to do is find the one master framework that is the one thing. What we are trying to do is find things we can do that may turn out to be hugely ahead of the curve that may turn out to be a really big deal, that may turn out to be the best money we've ever spent. I think long-termism definitely checks that box. And so, Open Philanthropy is never going to be an organization that is exclusively long-termist. But we do put a lot of money into it.

And the way you phrased it is one way of phrasing it. But if we also just phrase it another way and we say, why don't we try to focus our actions and our efforts on the things that might still matter a billion years from today, why don't we try to do the things that optimize for having the best future we can, for bequeathing the best thing we can to the future generations, for having the best overall story of humanity that we possibly can, having a healthy society, a society that makes good decisions, a society that's equitable and inclusive? I don't know. I don't think that sounds nearly so crazy. And I think if you were to walk around all day, this is something I've increasingly been trying to do myself because we've been doing division of labor at Open Philanthropy, and I've been focusing more on

long-termism. But if you just walk around all day, just thinking, well, what is it that's the best for helping the world be a good place a billion years from now? I think you end up doing a lot of really super reasonable things and maybe paying a lot of attention to things that should be getting more attention today.

ezra klein

Well, talk me through how even think of what the long-term in long-termism is because a critique we might have of just the way human beings are right now — I don't think we're typically doing things to make 24 hours from now the best it could possibly be, right? We're pretty far off of an ideal policy. So then, one version of long-termism is, let's just think about our children's world. And another is, let's think about 150 years from now, which is pretty far in the future. But I'd feel more confident predicting trends out 150 years, knowing I'll get some of them wrong. You then begin talking about a million years, a billion years.

holden karnofsky

A billion years, yeah.

ezra klein

What is long-termism to you? Because the further you go out, obviously, the more the uncertainty begins to bite. So how do you define it? And then how do you work within that uncertainty?

holden karnofsky

From a values perspective, I think that we should be caring about the whole future. But you raise an important point, which is the big obstacle to doing long-termist work is knowledge. And it's very hard to predict the future. It's very hard to identify actions that will matter that long from now. And so the vast majority of ideas we have, we probably will be wrong. And we'll probably be overconfident about what will actually matter a million years, a billion years, whatever.

So this is a huge challenge. And I think it's one of the downsides of being a long-termist and one of the reasons that I haven't put my whole life into it and never will. But that doesn't mean that we're totally helpless. It doesn't mean that we should just throw up our hands and say, let's just optimize for the next one year because that's the same as optimizing for the next billion years. And climate change is an example of that, where if you're always focused on the next year, you might say, let's burn more fossil fuels because that makes us all better off today.

But it's not some radical state of ignorance. It's not some giant unknown question whether climate change is going to make the long run future better or worse and whether it has the potential to make it a lot worse for a very long time. It's actually a real possibility. And so, I think as a long-termist, to do it well, you have to have taste and judgment. And you have to know when you don't know something, which is almost always, and when you might be on to something that actually could matter for that period of time. That's not an easy thing to do. And it's a pretty young idea. So I don't think we're nailing it, but I think someone should be trying it.

ezra klein

So I know that you started out in a place that's, frankly, more like where I started out or maybe even where I am on these questions, which is a bit more critical of the idea of long-termism, more critical of the appeal of long-termism within the philanthropic circles you run in. But recently, you've put a lot more emphasis in it. In Open Phil's 2018 update on cost prioritization, you wrote, "We'll probably recommend that a cluster of long-termist buckets collectively receive the largest allocation, at least 50 percent of all available capital."

Now, I know those numbers are subject to change. But it speaks to the fact that long-termism is an area you decide to place a lot more emphasis on in recent years. So tell me a bit about your trajectory on that. How did you become more persuaded there? And what did you become

persuaded of?

holden karnofsky

Sure. So as co-C.E.O. of Open Philanthropy, my job is to try to get ahead of the curve. My job is to look for ideas that are not only important but also neglected. And so, I'm always looking for what could be the next big thing that could matter for a ton of people that's not getting enough attention. And I deliberately seek out people and ideas that can introduce me to things that might do that. And so, through this, I have encountered the idea that this century, the 21st century, could be the most important century of all time for humanity. And a primary way that might come about is the development of A.I. systems that could cause a dramatic acceleration in science and technology, such that if you were to imagine a radical sci-fi future, a technologically advanced utopia, dystopia or anything between or even maybe a world that's not run by humans at all — that's run by AIs with their own non-human compatible objectives, which we can get to — a lot of people think that kind of long-run future is possible, but the right kind of A.I. could bring it very soon, could bring it this century. And once you think that you could be in that kind of century, now the timescales have collapsed. Instead of trying to make predictions about a billion years from now, we're trying to make predictions about the specific things that could happen in the next few decades that could matter for hundreds of years, thousands of years, billions of years. And so, when I first encountered this idea, I think it all just sounded too wild and too out there. And I really kind of mostly stuck to the work I was doing. But again, it's my job not to be too dismissive of ideas that could be extremely important and extremely neglected.

So, over the years, I've come to take it more seriously. And in particular, Open Philanthropy has had a team really focused for the last several years on taking this thesis about A.I. and the most important century and poking every angle at it, looking for the weak points, trying to figure out whether this is really plausible. And we've gotten to the point where I'm not going to say that this is something I know. I'm not going to say this is something that's going to happen. But I am going to say it's a serious possibility that I think deserves a lot more attention than it's getting. And the most recent kind of change for me is, I've been trying to get my thoughts straight. And so, I've started my own blog called Cold Takes, where I'm trying to lay out the case that we're in the most important century as simply as I can. And more generally, have been noticing that I've been taking on more unconventional views, more views that are important to what we're doing, but that are not widely held.

And I think it's important for me to be writing up those views in a clear and simple way in public, not only to help get clear in my own head about what I believe but also — because if I'm wrong, I want it to be easier for other people to encounter what I'm saying and to show me how I'm wrong. So this is a project that I'm on now, is taking these ideas that could be astronomically important that I've started to take quite seriously and continue poking at them by kind of putting them out there.

ezra klein

Well, let's talk about that project. We'll go deeper into these questions around A.I. in a minute. But first, I want to step back and talk about the big picture view of your— I always feel like you need drums and horns when you say this — Most Important Century series. One of the things you're really arguing there is that the future could be profoundly unlike the past in a way that it's true that 2021 is unlike 1900, but it's a lot like 1900.

It's still human beings running around. A lot of things are recognizable, a lot of the same religions. You had Democrats and Republicans in 1900. There's a lot happening then that was quite similar. But you're arguing here that 2300 could be basically unrecognizable as a world. So give me the basic case for that, the case for why the future might be wildly different than the past.

holden karnofsky

This is a big thing that has held me back from taking the Most Important Century idea seriously for many years, is that it just — if it's true, it implies we live in this wild, unusual time. And something I've learned is that if you just look at our time in full historical context, there's so many reasons other than A.I. to think that we do live in a wild time. So I think it's helpful to just kind of situate it in context.

First, there's the past. So the universe is more than 10 billion years old. Life on Earth is more than 3 billion years old. The whole thing of a species that's creating its own technology at all, even stone tools. That's millions of years old. So that's millions versus billions. That's the blink of an eye on galactic timescales. And then human history is also just very packed into the recent past.

So I think almost any metric you look at — economic growth, population, major technological milestones — more has happened in the past few hundred years than in the previous several hundred thousand or several million years. And that kind of points to this idea that there is a sort of acceleration that has occurred. Things have moved faster and faster. And if you simply project that acceleration forward and you say the acceleration continues or it's paused right now, but it comes back, in some ways, things going so crazy and the next few decades being more eventful than everything that came before is a continuation of a trend, not a breaking of it. Then, if you look to the future, I think there's other interesting observations about what a weird time we're in. Today's level of economic growth is a few percent a year. That doesn't feel that crazy to people. But it is not only an incredibly high level by historical standards, it's a level that doesn't look like it can go on forever. So if you try to project out thousands of years of growth of even 2 percent a year, it kind of looks like you've run out of atoms in the galaxy. You just can't do it. And so something has to change.

ezra klein

Hold there for a minute because I know the math on this, but it's very unintuitive. So, as I remember the calculation you run, 2 percent growth a year for 10,000 years will get you to a quantity that you basically can't run with the materials of the galaxy. But 2 percent doesn't seem that big.

holden karnofsky

Yeah, I mean, the specific quantity is at some multiple of the number of atoms in the galaxy, and there's very good reasons to think we would not be able to get even close to the edges of the galaxy in that time, because you're just constrained by the speed of light, if nothing else. So, yeah, 2 percent, I mean, it's exponential growth. And if you just plot it out 10,000 years, exponential growth tends to be very unintuitive.

Accelerating growth is even more unintuitive and even more explosive. And that is something we've seen accelerating growth or what's called super exponential growth. We've seen it in the past. And if we see it in the future, yeah, I mean, things go to the moon very quickly, and it's very unintuitive. But I do believe it's something that could happen.

ezra klein

But why this century? What makes this century so important?

holden karnofsky

So, three basic points. Point one is that the long-run future could be just radically unfamiliar. It could be a radical utopia, dystopia, anything in between. Point two is that the long-run future could become the near-term future if the right kind of A.I. is developed to accelerate science and technology dramatically. And point three is that that kind of A.I. looks more likely than not this century.

And then the bonus point four is that when you put the three together, a natural reaction is that this implies we're in a very special time. And it sounds too wild to be true. But point four is that if you step out and look at our place in history, it looks like we're in a very weird and wild time for many reasons that have nothing to do with A.I. And so we should be ready for anything.

ezra klein

I think actually just the idea of acceleration here is unintuitive. I think if you know the basic growth story, it's been fairly steady 2 percent growth globally for some time. People hear a lot about great stagnation. They hear about wage stagnation. It's been very hard for a lot of countries to break out of middle income traps. It doesn't intuitively feel like we are in an era of globally accelerating growth. So, walk me through the accelerationist argument.

holden karnofsky

Yeah, I think we're not in a period of globally accelerating growth. And I don't think we necessarily will be again. We could just end up with a world that stagnates, where growth slows, like you're saying. The case for accelerationism would be — so the basic idea is, if you look at most standard economic growth models, there's this potential for a feedback loop. This is something that can happen. It's not something I think is happening today. But you could have a feedback loop where every time you get more resources, more food, whatever, that leads to more people. More people have more ideas. More ideas leads to innovation, and therefore, more resources. So you get resources, people, ideas, resources, people, ideas, and you get a feedback loop. And that causes accelerating growth that can be this very explosive dynamic.

And it looks reasonably likely that this has described periods of economic history. The people refer to the Malthusian dynamic, where people didn't get much richer, even though they were improving productivity, because the additional resources would just go into more people. And so, you see that pattern.

And what happened, what changed a couple hundred years ago or so is called the demographic transition where it stopped being the case that more resources meant more people. And now, of course, when people have a lot of wealth or have a lot of resources, they tend to just be richer. It doesn't cause them to have more children. And so —

ezra klein

It does the opposite in fact. They have fewer children.

holden karnofsky

Exactly. And so if that's the dynamic we're in, then, yeah, you would, by default, expect that we will get stagnation. And I think that's a serious possibility that our long-run future looks like economic growth has to slow dramatically. I think there's a lot of reasons to think that is our default. The question is, is there a way to get that same function provided by people, provided by something else, such as A.I., that can be straightforwardly produced by more resources.

So today, we have A.I. systems that are cool. They can beat people at chess. They can transcribe audio. What they're not doing is they're not doing that innovation part of the feedback loop. They're not creating new technologies autonomously. They're not advancing science and technology on their own because there's such limits to what they could do. And the question is, if that changes, if AI is developed that could be as good as humans — doesn't have to be better — at sort of pushing science and technology forward, then you get the feedback loopback back. And then you could get an explosion.

ezra klein

So there are a number of things I want to question in this or describe more. But I want to first start at what I think will be the natural objection, which is something you didn't spend time on there. In a population-resources-ideas feedback loop, resources is a part of this. So I think it's gotten the worst name recently.

There is a broadly held view — I hold it — we talked about climate change already in the show— that we've had a lot of growth that is using up a lot of resources in a way that, for all the good it's done, has also done a lot of harm. It's put us in a very dangerous climate situation. It has led to a tremendous amount of extinction of other species.

The idea that you would just accelerate growth from there, well, with what resources? Oftentimes, growth does not create resources. It consumes them, right? We're consuming fossil fuels that are finite on this planet, et cetera. How do you think about the resource constraint? Or do you not see it as a constraint? Talk about that piece before we go into what would happen if you blew up the rest of the loop.

holden karnofsky

First, I just want to be clear that I'm not talking about this possibility as this is exciting, and we got to do it. I'm talking about it as this is something that might happen, for better or worse. And I think it could be very good or very bad. But in terms of resources, so humans right now are the only thing that sort of have ideas, that sort of create new technologies that advance science. And humans have a lot of needs.

And there's a lot of resources for humans that are really hard to replace. There's only one planet that we really know of where it seems pretty doable for humans to exist for a long time. And once you have a different kind of technology fulfilling that part of the feedback loop, that kind of technology does not need all the same resources that humans have. What do you need in order to run more A.I.s? You need more computers. Well, you need some things for that. You need metal. You need electricity. You need cooling. But there's not that much that you need to run more computers. And actually, all the things you need are very abundant in space. And all the things you need to get to space, I think, can be built with those sort of limited set of resources. So I think that's kind of the whole idea, is that it's a lot easier to build more A.I.s than it is to build more humans. And so this loop could quickly get out of control if that dynamic changes.

ezra klein

I want to put a pin in there's a lot of metal in space because I think it's actually an important part of your vision, a lot of other people's visions that we should come back to in a second. But in terms of here, this is always a question I have about the AI and locking tremendous economic growth question, which is, oftentimes, the constraint on an idea improving people's lives is material in some way or another.

So, for instance, there is a lot we could do with better analysis leading to better drug innovation. But actually, a huge constraint in drug innovation is you need to run a lot of trials on human beings. You need to actually test things in the real world. It slows everything down. It's a big deal. It's not clear to me, in the first approximation, where A.I. helps on that or how much it actually unlocks.

Or in terms of things people might want, houses take wood. Cars need to be built. It is true that the A.I.s themselves would not need a lot of resources. But in order to get a huge amount of pressure on the growth number, which is measuring things the economy actually produces and people consume, a lot of this would have to be built. It's not going to all be digital t-shirts in the metaverse.

And so, how does A.I. create more resources, as opposed to even potentially disastrously using them up? I mean, should I look at resources as fixed, and more growth will just consume them faster? Or should I look at them as like a pie you can expand? How do you think about that part of the equation?

holden karnofsky

I think my first answer to the question is that if you assume that you can't translate these A.I. resources into human resources, it doesn't change the fact that we're looking at an enormously consequential development. So if you have this kind of dramatic feedback loop and a dramatic increase in sort of the reach of our world, of our planet's artifacts, if there could be sort of this growing population of A.I.s that is expanding throughout space, that's a very high stakes situation for humans. If it turns out there's no way to convert all of that wealth into wealth that makes humans' lives better, well, that's really bad news. It doesn't mean that this is a nothing burger, though. It kind of might mean this is really bad news. So when I talk about a giant acceleration in science and technological advancement, that doesn't have to be an acceleration in medicine. But it's an acceleration in some sort of general potency, some sort of resources, some sort of ability to make something happen. And the less compatible that is with humans, the worse. Then it sort of becomes up to us.

I mean, if this happened, could we harness it in a way that it was to humans' benefit? Could we take a very large supply of sort of digital minds or A.I.s having ideas and use that to make the world better? Or is it just going to turn into this sort of runaway train? And I think that's a question for us.

ezra klein

Well, give me a concrete example. What's an example of a way, if we unlocked this boundary we have on creativity and innovation, right? You have 5 billion A.I. minds coming up with cool ideas all the time. When you think about how that could lead growth to go vertical, what are the kinds of things you're thinking might emerge? What problems that we have might they solve? I mean, it's all fine to talk about growth and GDP and A.I.s, but paint the sci-fi utopia or dystopia for me. Let's get tangible.

holden karnofsky

Sure, I'll start a little bit more tangible, and then I'll get a little bit more out there. So for the little more tangible, we can talk about energy. Energy is definitely something that if you had a lot of minds and you had a lot of resources, you should be able to get a lot of energy. You should be able to get it very cheap. And you should be able to get it in a way that isn't necessarily involving any greenhouse gas emissions or anything like that because there's just plenty of energy to be had if you had big enough and good enough solar panels. That's something that I think could absolutely come out of this loop.

Then if we talk about health, I mean, you might be right that health is always going to be bottlenecked by human trials. But if you're able to do enough simulations, if you're able to have enough minds on the problem, you really could come up with a very large number of candidate drugs at once. You might have to wait a few years for the clinical trials. But it's sort of the sky's the limit in terms of how much you could improve health.

And so, that's the answer number one, is health and energy. Well, I mean, that's an awful lot. We could go through other parts of the economy, but dramatic changes in health, lifespan and energy would certainly be a big deal.

Now I'm going to go more to the wilder end of the spectrum and the more speculative end because I think it's important for people to not have their aperture too narrow and not be stuck on things that feel like today, when we could be looking at dramatic changes. So I do want to talk about the more radical end of things. It's very important not to have too narrow an aperture. It's very important not to insist on thinking about things that seem real today when we could be confronting a dramatic acceleration.

So, in the Most Important Century series, I try to just give one concrete imaginable example that won't necessarily be the right one of how technology could get super crazy advanced and give us a huge amount of control over our environment. I call the example digital people. This is very related to the idea of mind uploads.

And it's the idea that you could have sort of people made of software, fully conscious beings. They don't have to be just like us, but they would be something that we decided were properly referred to as our descendants, some sort of digital mind that we would look at and say, I care about that digital mind. That digital mind matters. And identify with them. Maybe I think of them as part of our civilization.

And just from that, just from having either a digital running of a conscious human mind — and we could talk about whether such a thing would be conscious or what that would even mean, happy to get to that if you want — or just something else that we identify with for any reason, that could get extremely radical because any kind of digital being — it could be in a virtual environment, which means that it experiences anything that the runners of the environment want it to experience — that can be unlimited good. That could be unlimited bad. There's no reason that digital people need to have any kind of health problems or any kind of lifespan limits.

And I'm not saying this is necessarily a good thing. There are very scary things about the idea of digital people. You can imagine a world where people are able to lock in certain values to their society forever because of the enormous amount of control you would have over those virtual environments.

So I think digital people are something where if you just imagine it's possible, you visualize it to be concrete, you draw out the implications, it's very easy to get very quickly to a utopia or a dystopia that just goes all across the galaxy and lasts forever. Now, is that literally what's going to happen? No, it's an example of what could happen if technology advances enough. And there's probably many more I haven't thought of.

ezra klein

I understand the idea of AI.

holden karnofsky

Yeah, sure.

ezra klein

Digital people is a slightly different concept.

holden karnofsky

Yeah, it is.

ezra klein

What makes you think that that is something that we could do at a level of fidelity that we should be thinking of them as conscious, having rights, people we could simulate on? You have an interesting idea that one way digital people could really help is by accelerating social science. You could see what happens if you make bunches that meditate two hours a day every day, and others don't. But that you could have the fidelity in them that a simulation of a digital people would be relevant to, I guess, we'd call them real people, physical people. Give me the plausibility argument.

holden karnofsky

I mean, there's many ways you might imagine having digital beings that we cared about for one reason or another, but the plausibility argument I talk about, just the most concrete, is just, what if we could simulate everything that's going on in a brain? And brains are not these enormous rare artifacts. There's a lot of them.

And there just doesn't seem to be a particularly good reason to imagine that we wouldn't be able to find some way of simulating a brain so that it is performing all the same important operations, all the same information processing, turning the same inputs into the same outputs, especially if you're

imagining a dramatic acceleration in the population of A.I.s, which could be doing the science, which could be understanding how to run these simulations, which could be providing the computers that we've talked about. You could have as many as you want.

This is not something where I've decided, oh, yes, this is definitely doable. This is just something that's more like, I don't see why not. I think if we were in this limit of having sort of infinite computing resources, I would expect this to happen by default, but maybe it won't. But it's definitely an example of a way the world could go very wacky very quickly.

ezra klein

This loops us back to that economic feedback idea because this is something I think is really important in your thought. And it is worth a sitting for a minute, which is that you really identify the core constraint on growth as population. And I don't think that's where people tend to think about it. They think of resources. They think about ideas.

But in a lot of these different ideas, whether it's A.I. — and it does not have to be sentient; it can just be idea-generating A.I. that never has a feeling of its own, whether it's digital people, simulations of us, whatever it is, you're imagining a world where you just completely blow the top off of the population constraint. So tell me a bit about the population constraint as it operates in your thinking. Why do you see that as what would unlock so much potential for humanity?

holden karnofsky

. This is largely coming from the different models that economists are using to theoretically model growth. And they all have this kind of point of commonality. Not all of them, but it's just very common. And Open Philanthropy, we did a report called "Report on Explosive Growth" where we kind of very broadly surveyed the economic literature on how people model growth. And there's a lot of ways of doing it.

But this feedback loop is just a very common property. It's a very common theme. And it's true that I think most people don't tend to think of population as this essential input into growth. But I think that comes back to long-term, short-term again. A cool paper that I recently wrote about, there's a paper by Chad Jones that just talks about what you think about long-run growth if you're using what he calls the semi endogenous growth models.

And it does say that over the last 100 years, a lot of our growth has not been about population. It has not been about population growth. It's been about increasing education. It's been about reallocation of resources. But these things, they have limits to how much more of them we can do. In the long-run, it's kind of hard to see how you get continuing high growth or continuing improvement in living standards per capita if you're not having a rising population.

So I think the idea of population is that humans are the only things that produce ideas, innovations, new technologies. You need them to do that. You need more of them if you want more of that. If ideas are getting harder to find over time, then you need even more of them over time. And so, this becomes this very important factor.

[MUSIC PLAYING]

ezra klein

So I do think taking the dystopic possibilities of all this seriously is important. Something the sci-fi writer Ted Chiang once said to me has stuck in my mind when I asked him about if he thought we'd ever create a sentient A.I. And he said, no, he didn't really think so, but maybe we could. But we shouldn't because far before we ever created sentient A.I. that could kill us all, we would create A.I. that we could make suffer dramatically.

When I hear about this possibility of creating these virtual environments, whether or not it is digital people or just us people, our people, I've always said that the most likely dystopia in the next couple hundred years always feels to me like the Ready Player One dystopia, where for vast amounts of the population, the actual opportunities in the world have just become so diminished that they are in a Brave New World-like V.R. rig. And things have just become artificial in a thin way as a way of escaping from a grim reality. When you say you're not sure if this is good or bad, I think that the implication when we talk about growth is that it's all going to be great. What are the things that worry you here?

holden karnofsky

Things that worry me, number one would be misaligned A.I.s. So I take the possibility seriously. And I know it sounds wacky and like a sci-fi movie, but I take the possibility seriously that things could just spin out of control. You've got A.I.s that can grow the population in a much faster way than humans. And you've got A.I.s just running the world, setting up whatever they're setting up over the galaxy, based on their own sort of random objectives that came in because they weren't carefully designed and are not compatible with humans.

So that seems like probably the thing that I feel best about saying, let's really make sure we don't get that. That just doesn't seem like a human-friendly world in any way, shape, or form. Then, what are other things I worry about? I mean, in general, I don't like the idea of rushing headlong into this world, of having a world where all the people who have inclusive and humanitarian values are focused on other issues.

Everyone working on A.I. is just someone who loves doing fun, sexy stuff and pushing tech forward. Then we kind of rush headlong into this world where it just becomes a race. And it just becomes — you have this dramatic acceleration. And the people who are the least careful and the most eager to win the race end up running the world. And those are not the kind of people that I hope will end up running the world.

So, those are two things I worry about a lot. And a lot of what feels central to me when I think about what do I hope for, if this is the most important century, is, I hope we find a way to have more reflection than either of those implies. There's this term Toby Ord uses in the book, "The Precipice," called the long reflection. And it's the idea of having civilization take its time, in a sense. Use technology to give itself time to improve and to have moral progress and to not build some incredibly large, powerful, galaxy-spanning civilization before it's something that can be done in a fair, inclusive, thoughtful way.

ezra klein

So, one of the ideas here that you talk about is also that we're in what I've come to think of as a quarter of existential risk. That until the 20th century, humanity hadn't really created technologies that could possibly wipe out all of humanity without a really tremendous amount of effort, right? You really would have needed a lot of people going around, individually killing everybody else.

But then we created nuclear weapons, and all of a sudden, we could. We've talked a little bit about the possibility of synthetic bioweapons. Misaligned A.I. is something people worry about on this score. This is a big point of Toby Ord's book. We are one of the first generations to really have the power to end human life. And we're not that wise. Our technology has advanced far quicker than our governing institutions, far quicker than our moral intuitions, I would say.

And so, a point you make is that if you get to the other side of this technological era, if you can diversify humanity into other worlds, into other formats of being alive, like computers, maybe there isn't quite as much existential risk anymore. But between here and there, there's a lot. That also strikes me as a reason this is an important period because we may be more protected from wiping out our civilization in a couple hundred years, but we're not now.

holden karnofsky

Yeah, that all sounds right to me. There is some future state where we've likely become very technologically advanced, where there's not a lot more advances around to kind of shake things up, where there is a great amount of control over the environment, where we're able to be spread throughout the galaxy, have the stability that comes with that, where we might have the stability that comes with not having aging or dying be necessary or just having those take a very long time.

So there's a lot of reasons I think that you might expect just the world to reach a state at some later date where however things are, that's how they're going to stay, at least for very long periods of time. And I think that could be a very good thing or a very bad thing. It just depends on what that state is.

And that's what gives me so much vertigo about the idea of the most important century.

It's just, things speed up a lot. And you get to some sort of radical future. And the seed of it, the starting point, seems like it matters an awful lot. And it's this huge opportunity for things to go very well or very poorly. And I don't feel like there's a lot of seriousness about what to do with that situation.

ezra klein

So should we maybe just not? Should we maybe just not be trying to build A.I., maybe just not be trying to push some of these technologies forward? Should we maybe just not be trying to grow as much as we try to grow?

holden karnofsky

I think if it were up to me, we would be really slow and deliberate and reflective about it and sort of have some process for making sure that a lot of people were included in a serious reflection on whether this is a good idea before anything really dramatic and irreversible happens. It's not up to me. I think this is coming. I think the economic incentives to create this kind of A.I. are very strong. The idea of having the whole world just not work on it doesn't seem viable.

I also think there are tremendous potential benefits from A.I. and from all technology. Like I've said, I think that the pattern of economic growth seeming like a good thing does look true over the last couple hundred years and probably is true over the coming decades, at least. So I just don't know that it's realistic to go the way you would kind of go in a happy dream world. And I think instead, we have to take this as something that's coming our way and think about what to do about it.

ezra klein

Does the entire vision for you actually rely here on this kind of A.I.? Which is to say that if, for some reason, it just turned out that you couldn't create this kind of transformational A.I., that it was never able to learn on enough real world data to create real world innovations, that the computing was too difficult — I don't know — something was found, and it just doesn't — there's a limit we hit and we can't go beyond that on creating computer intelligence, does then the future just look pretty normal to you?

holden karnofsky

It looks a lot more normal. It would definitely be a big change in my views. It would remain the case that we live in a very strange time. And I feel like we should be ready for wild stuff to happen next. It would remain the case that anything that kind of went into that feedback loop or dramatically accelerated science and technology for any other reason could be enormously high stakes.

There are other technologies I could name that I would consider to be still extremely important, still extremely high stakes. I think the climate change situation would be one of the biggest things looming in sort of this could make a permanent difference kind of zone. So would the risk of pandemics. So I think there'd still be a lot to think about. I think we would still live in an extraordinarily strange time.

But learning that there's no kind of digital mind that could do what humans do to advance science and technology, including a simulation of humans, a digital simulation of humans learning — that's just off the table — that would dramatically move me toward expecting stagnation and thinking about what to do about that.

ezra klein

So let's live in that world for a minute. Let's not think about digital people, A.I. minds, none of that. Something you tapped earlier is the possibility of easing the resource constraint, because you are mining the incredibly abundant minerals and metals and asteroids and in other parts of the galaxy. Even before you get to settling all over the galaxy, you actually could get a lot of resources out of things that are just flying around space. How important is that as a set of technologies for opening the resource constraint, as opposed to the population one?

holden karnofsky

That's not something I have really strong views about. I think there's clearly resources that are important for us that we're not going to be able to just get that way. Having a habitable planet and a friendly climate to humans, you can't just take a bunch of metal and somehow make a friendly climate out of that. So I do think eventually, it should be possible to get all our energy from clean sources.

But I think there is a real possibility that if this A.I. thing turns out not to be feasible, we could be looking at a world where we really do stagnate in terms of living standards, and we don't have as much of a growing positive-sum economy anymore. And that becomes a different kind of challenge.

ezra klein

Would that be a bad thing?

holden karnofsky

I would think by default, yes. It seems like you've got kind of a wind at your back if there's more to go around every year, more wealth, more just things that make people's lives better. And there is still an enormous amount of poverty today that is just really severe and is quite material. So it seems like there's still quite a lot of work to do in reducing poverty. It's a lot of what Open Philanthropy focuses on. It's what I've spent most of my career on. And I think it's really important.

So, yeah, I think the idea that you stagnate and the idea that you just have what you have, I hope it would happen after there's no more poverty of the kind that we see in lots of the world today. But even if that happens, I think it makes me nervous to have a world where there's no wind at your back of that kind. And it just, we have what we have, and we're all just fighting about who gets the most of it.

ezra klein

Well, let me take the other side of the argument for it to be provocative.

holden karnofsky

Yeah, sure.

ezra klein

So when I talk to you or anyone else who believes in something like the most important century, I hear a lot of, I'm not saying this is good or bad. I hear a lot of, we could invent misaligned A.I. that wipes out all of humanity. I hear a lot of, maybe we'll create digital people or at least some kind of morally important, morally consequential digital life form. And I look around, and I look at how we treat animals. And I think, well, that may not go well at all.

holden karnofsky

Agree.

ezra klein

I look at the creation, of course, of weaponry, which you've not talked that much about here, but that could be unbelievably destructive. And there's, of course, an alternative vision, which is, I had Richard Powers, the novelist, on the show a little bit back, who wrote Overstory, just wrote the beautiful book "Bewilderment." And I mean, his view of where humanity should go is not non-technological, but it is into more of a relationship with nature.

I think we called it, in a nutshell, a scientific animism, that the world is getting richer, but we don't need to be that much richer for there to be a lot to go around. We have the capacity to, over the next 30 or 40 or 50 years, power things through solar if we chose to. I mean, the same constraints on our wisdom that keep us from responding to climate change well now, they bite as hard or harder in the kinds of futures you're talking about.

So, maybe the future we should be working towards is not this unbelievably high risk gamble of spreading out throughout the stars and maybe enslaving A.I. or being enslaved by it, but just trying to chill a little bit, share a bit more, protect the climate. Why should that not be the vision?

holden karnofsky

One, my answer is it doesn't seem to be an option. I mean, we try to focus on things we can do today that will be helpful for the world we're actually heading toward. And if this kind of A.I. being a giant driver of acceleration happens to come out, I mean, I don't think there's much I can do, sitting here today, to just decide that's not the way we're going to go. In the other world, I mean, I think that vision kind of sounds nice at a high level, but there's just an awful lot of horrible grinding poverty in the world today.

And so I don't think that would be a very appealing vision until people have a basic standard of health care and a standard of living and good nutrition, which a lot of people in the world just still don't have enough calories, are still kind of stunted because they didn't get enough to eat as kids. So I think that's very important. There's a lot of work to do there. And that's not something that's going to happen if we don't see more economic growth.

I also think there is — I don't know, but I think the idea of zero economic growth just makes me nervous. I think it should make anyone nervous. Because as long as we have economic growth, there's at least things that are happening in the world that are positive sum. There are projects you can embark on, where you can say, I'm doing this to make everyone's lives better.

And at the point where the only change in resources is zero sum, the only thing that helps one person hurts another person, that's a different dynamic than we've had over the last few hundred years, during which, as I said, it looks like life has been improving. So it's at least something that makes me nervous. It's just, I don't know what the world looks like in that kind of situation. Maybe it turns out wonderful. Maybe we all learn to just have perfect equanimity and live together in harmony and appreciate what we have. But it does make me nervous.

ezra klein

But I want to pick on this for another minute because I do think it gets to a constraint people will take seriously in a low growth world and then don't take that seriously in the hypergrowth world — say they do, but don't — which is you were just saying that there are many, many people who don't get enough calories around the world. We could change that.

holden karnofsky

Sure.

ezra klein

We could.

holden karnofsky

Agree.

ezra klein

We could change our energy systems. Many of the things that make the world we live in at its current state way worse for billions of people than it needs to be are solvable, and we don't solve them. I mean, not only do we not solve them, we are not going to get close to vaccinating lower income countries this year. We could. This needed a couple billion bucks and a good global mobilization. And we're not going to do it, even though we could absolutely afford to do it.

And the same shortsightedness, the same lack of compassion, the same inability — actually, I'd say, an expanded inability— as hard as it might be to see somebody in another country as kin, it will be that much harder to see a digital something rather as kin. Hard as it is to see a cow as kin, again, I think the way we will treat non-player characters in this world could be really scary. And so, if we haven't been able to make something closer to a paradise out of this world, I'm not saying why should I believe we will make the hypergrowth world better. I'm more saying that why should I want to give us a chance to make it as bad as we could.

holden karnofsky

I mean, first off, I would say that we haven't solved the problems that exist today, but we've made a lot of progress on them. And this is where I agree with the news floating around about how the last few hundred years look for humanity. They look good. It looks like there's been a lot of progress on a lot of things. So, there has been progress. And we do get somewhere. I mean, we don't get immediately to where I wish we were as fast as I think we could. But I'm not sure exactly what you're asking here. Are you asking about why the hyper —

ezra klein

I'm making you — well, I'm trying to just make you question some assumptions.

holden karnofsky

Which assumptions do you want me to question?

ezra klein

Simply the assumption that if you want to say you don't know if all this will be good or bad, maybe the possibility that it will be bad makes it such that we should be stopping people from trying to do it. And that if we can even get the world we have right, a world with this kind of dizzying change may be beyond human capacity to manage in any kind of even basically decent way. And so we should fear it as more likely a dystopia than a utopia or even something in the middle.

holden karnofsky

Yeah, I basically am just in agreement as a maybe — I mean, as this is a possibility. I mean, I'm just in agreement with you. I'm not sure I have an assumption that contradicts anything you said. I think where I am sitting is I'm saying, do I want my role to be trying to stop people from building powerful technologies to address real problems that exist in the world today because of this risk it might accelerate? I mean, I don't want my role to be that, partly because it would be a strange thing to be doing, based on some grand vision for how the long-run future is going to go that I don't think I have. And I do think we have short-term problems that are worth addressing. But also, I don't know — I don't think that's a path that looks very promising to try and go down. And the three criteria for Open Philanthropy's work are importance, neglectedness and tractability. So importance is how big a deal is an issue, how many people does it affect. Neglectedness is how much attention is it getting. We'd rather work on issues that get less attention. And tractability is what can we do about it. Do we see a

path to victory? Are we going to matter? Open Philanthropy cares about all three of those. So when I think about the possibility of the most important century, I'm looking for paths and things we can do that will make the best of it.

ezra klein

So let's say you have somebody listening, and you buy into this, or at least, you buy into it as a real possibility. And you want to try to help make the best of it. What is tractable for an individual here? What does this imply for just somebody's life, if they want to try to live on this time scale?

holden karnofsky

The first answer I want to give is that I wish I had a better answer. I wish I could just tell you, hey, I've not only figured out what's going to happen, which I haven't — I figured out what needs more attention. But I figured out exactly what to do about it. But I haven't done that either. And one of the things I say in my blog post series is, if you're thinking, hey, this could be a billion-dollar company, maybe the right reaction is, yeah, awesome, let's go for it.

And if this could be the most important century, I think my reaction tends to be just like, ooh. I don't know what to do with this. I need to sit down. I need to think about it. I think there's a number of reasons that it's actually very hard to know what to do. And I think we need more attention and more thought and more reflection because most things that I could name as an attempt to make the most important century better could be really good or really bad, depending on your model of what the most important considerations are.

So I can get to that in a second, but I don't want to be totally gloomy about it. I think there are things that look robustly helpful. There are things that look good. One of them is A.I. alignment research. So just, if we could get to a point where we're capable of being confident that we can build advanced A.I. systems that aren't just going to run the world according to whatever objectives they have, and that is a field that exists. That's a field that people are working on. We fund work in it.

Another robustly helpful thing is just trying to find and empower more people who are seeing the situation as it is, taking it seriously, being as thoughtful about it as they can be, and approaching it from the perspective of what's best for humanity, instead of what's best for them narrowly. So I do think there are some activities. And there's some discussion of it in the blog post series.

ezra klein

And then finally, always our last question, what are three books that have influenced you you'd recommend to the audience?

holden karnofsky

First book is called "Due Diligence" by David Roodman. He's my coworker now, although he wasn't at the time he wrote the book. It's about microfinance, which is very small loans to low-income people. And there was this massive debate about how much microfinance helps people, whether it's the best way to help people.

And "Due Diligence" is an attempt to sort through a very overwhelming debate among experts and academics from looking at it from multiple angles, getting the right answer. I think it's a great book to read about how to think. I think it's better for that than a lot of books about how to think, although I generally think writing is a better way to learn critical thinking than reading anything. I think this is one of the best things to read for that.

Book two is called "The Lifeways of Hunter-Gatherers" by Robert Kelly. I want to be upfront. It has a lot of really dry theoretical parts. There's many parts I skipped entirely. There's many parts that I've read maybe a dozen times by now. It's trying to pull together all the evidence we have about what life is like for people who live in a hunter-gatherer lifestyle, which is often assumed to represent what the distant past of humanity looks like, before the Neolithic Revolution 10,000 years or so ago. And it's

interesting to see how someone tries to piece together the clues about our very distant past. And like I've said, I've generally been obsessed about thinking about long time frames and getting my head into that space.

Third book is about the future. I'll recommend "The Precipice" by Toby Ord, which we've talked about. It's about the idea of existential risk. It goes through the different things that might happen that could permanently cut off humanity's future. I think it's a really good read. And then I want to throw in two bonus books that haven't been published yet.

ezra klein

Do it.

holden karnofsky

So one of them is "What We Owe the Future" by the Scottish philosopher Will MacAskill. It comes out, I think, in September 2022. And it's about the idea of long-termism. It's about what are the things we could do today that could matter to all the future generations. And shouldn't we care about that, when those future generations don't have a voice in the things we're doing today that affect them?

And the final book does not have a title yet. I can't promise it will ever exist. But it's by my wife, Daniela Amodei. And I would expect it in a few years. And it's about the history of people using science and technology to get better gender equality, so things like the pill, things like formula feeding, so that the whole burden of childcare doesn't fall on the woman, things like I.V.F. for more options in reproductive timing, and about how in the future, it's going to be necessary to use more science and technology and to have better science and technology to achieve better gender equality.

ezra klein

Holden Karnofsky, your blog is Cold Takes, which we will put a link to in the show notes. Thank you very much.

holden karnofsky

Thank you. [MUSIC PLAYING]

ezra klein

"The Ezra Klein Show" is a production of New York Times Opinion. It is produced by Jeff Geld, Rogé Karma, and Annie Galvin; fact-checked by Michelle Harris; original music by Isaac Jones; and mixing by Jeff Geld.

[MUSIC PLAYING]

Listen · 1 hr 30 min

[MUSIC PLAYING]

EZRA KLEIN: I'm Ezra Klein, and this is "The Ezra Klein Show."

Over the past — I don't know — five, six years, I've been very influenced by the effective altruism movement. On one level, effective altruism is simple. It asks, how do we do the most good we can with the money and the resources we have? That turns out to be, one, a deceptively difficult question and, two, weirdly, one that we don't ask all that often, one that oftentimes you think people are asking and they are not.

But the difficult parts are maybe more interesting. How do you measure the most good? What about when you think something is good, but it cannot really be measured? Who defines good? Who verifies impact? How do you judge the value of, say, supporting art against the value of building housing for the poor?

Effective altruism has roots in the academy. Philosophers like Toby Ord and Will MacAskill and Peter Singer, they've been central in creating the movement. And importantly, they're central in the way the movement thinks and reasons. The culture of effective altruism, in my experience — and this is both its best and worst quality, in a way — can feel like a philosophy grad seminar that never ends.

By that, I mean it delights in taking the logic of its questions as far as it will go. It's unafraid, even ecstatic, to follow answers that strike others as very strange or unintuitive, sometimes even cruel. It's always, always questioning its own assumptions and everyone else's. It can, in my view, sometimes be performatively cold or logical in a way that's actually quite narrow about human flourishing. But as I've said, I have learned a lot from these thinkers.

What's interesting to me is that effective altruists tend to come out on two very different ends of what you might end up worrying about. Some take the analysis and end up worrying a lot about very provable, very, very well defined interventions, like distributing malarial bed nets, because those are the interventions we have studied with the best research. We've done randomized controlled trials. And we know they save lives cheaply.

And then others come out obsessed with much more speculative threats, like, say, killer artificial intelligence because if you start running the thought experiments, if you assign a moral weight to the future and all the potential human beings in the future, anything that could end humanity tomorrow really demands our attention today. If you can improve the likelihood of countless generations even a little bit, that really nets out to a big impact on the future.

No one represents the dual and sometimes warring impulses of effective altruism better, to me, than Holden Karnofsky. Holden was one of the co-founders of GiveWell, which measures the effectiveness of different charities and recommends the ones it is most confident can save lives cheaply. But then he spun out of that to found Open Philanthropy, which is more on the speculative side of effective altruism, the thought experiment side.

But because he's actually giving away money and making grants and shaping sectors and commissioning research, he and his team have to be pretty serious about how they approach questions that are not always considered seriously. They don't even always sound serious when they are spoken aloud. So there is a practicality, even a groundedness, to his analysis, even when it ends up in very, very strange places.

And more recently, Holden has been going to stranger places still. He launched a blog called Cold Takes. And on it, he's making the case, in piece after piece, that we live in maybe the most important century humanity will ever have, the most important we have had or will have, that the future could be wildly unlike the past. It is real mind expanding stuff.

And this episode — warning — will go to some very strange places. This is not our normal fare. But I'd urge you to follow it there. There's a concept Holden and I talk about here — worldview diversification, the practice of recognizing that we're not always sure how we should look at the present or the future. And we should keep a number of different possibilities, some of them even outlandish, alive in our minds. We should even keep them alive in our actions, rather than trying, as we so often do, to choose between them.

This is something I try to do on the show. You can think of each episode as a little exercise in worldview diversification. And I try, as you know, to stay away from the choosing, except when we really have to. But it's particularly something we do in this episode. Is Holden right about some of the more wild predictions we discuss at the end? I don't know. On some level, I have a lot of trouble believing it, but I do think it's good to stretch the predictions you're willing to entertain, if only to help you see the present more clearly. As always, my email: ezrakleinshow@nytimes.com.

[MUSIC PLAYING]

Holden Karnofsky, welcome to the show.

HOLDEN KARNOFSKY: Thanks for having me.

EZRA KLEIN: I've wanted to do this a long time, man.

HOLDEN KARNOFSKY: Cool.

EZRA KLEIN: It's a minute in the making. All right, I want to begin at the beginning for you, before Cold Takes, before Open Phil, back in the GiveWell days. Tell me the story of why you started GiveWell.

HOLDEN KARNOFSKY: Sure. So I was a few years out of college, and I wanted to give to charity. And I had the immediate thought that maybe I could find a website that would just tell me where to give to charity to sort of get the best deal. I think you could maybe think of what I wanted as a Wirecutter for charities, although there was no Wirecutter then, either. What I wanted was to kind of help as many people as I could with the money that I was giving. And I tried to find this, and I couldn't.

And my co-worker, Elie Hassenfeld, and I were in the same boat on this. We were going through the same journey together. And we both decided to create the website that we wish existed, which is GiveWell, which is just public recommendations, telling people which charities can help their money help the most people possible.

EZRA KLEIN: So I looked for things like this in this period. And I found things that at least purported to be this. People might be familiar with Charity Navigator. There's a lot of cross-charity budget comparisons. Why weren't those what you were looking for?

HOLDEN KARNOFSKY: Sure, so at the time we were doing this was around 2006, 2007. And the obsession in the charity evaluation world was around the overhead ratio, or how much of a charity's budget does it spend on the so-called programs versus the so-called overhead. This is an appealing thing in that you have to report these numbers to the I.R.S., so you can get this number for any charity. But it really doesn't seem all that relevant to me.

For example, if you have a charity and you decide to pay your top talent more or you get better information technology, and now you're doing way more good and helping way more people, is that overhead? Is that somehow a waste of money? Is that, somehow, you did the wrong thing? And so, we wanted something different. What we wanted was this question of, what charity is going to help the most people per dollar that I spend?

And a crucial difference is that we were not in a place where we had a charity we already liked, and we wanted to look it up and make sure it was legit or that it wasn't fraudulent. We wanted something different. We wanted recommendations of which charities out of all the many you could do the most good. And so kind of like how the Wirecutter will actually kind of tell you what five products to maybe buy, instead of you look up a product and they tell you if it's going to break.

EZRA KLEIN: On a note, the Wirecutter is now a New York Times Company, and I've not paid Holden to make all these Wirecutter analogies. But getting at this, one of the reasons there was an interest in things like overhead spending is it is cross-comparable across charities. As you mentioned, that's a number reported to the I.R.S.

There are a lot of charities. They do very different things. Some of them work on international poverty. Some of them work on public health. Some of them give money to ballet. How do you begin to narrow down and test which ones do the most good for the least money?

HOLDEN KARNOFSKY: Yeah, so we started with some criteria. We wanted to find charities that were proven, cost-effective and scalable. Proven means that there's strong evidence that what they're doing helps people. Cost-effective means that they're helping a lot of people per dollar. And scalable means that there's room for more funding. It means that if you give more money, more people will be helped. So we're not just talking about a charity that did something great once and wants more money now.

Those criteria will not capture everything good. And I imagine we'll get to this, that I've later branched out into other kinds of giving that don't have necessarily all those criteria. But when you have those criteria, if you look at any charity and you ask whether it's really helping people, you've got an academic literature to review. Does education interventions, do those

help children learn better? Do health interventions help people live longer? And what you can do is you can start looking for the programs that are the most proven and the most cost effective. And then you can find charities that do them.

So that's a way to narrow the field. And one of the things that we learned early in GiveWell, which is a lesson that's kind of stuck with me throughout my journey to where I am now, is that if you want to do the most good, in some ways, the worst place you can start is your neighborhood, your friends, your country, because if you look at the whole population, in this case, with GiveWell, there's just a lot of people living in countries that are extremely poor by U.S. standards, and your money can go a lot further there. You can help a lot more people with the same money if you're willing to broaden your moral circle and include more people in it.

EZRA KLEIN: Talk me through this tangibly. What is the first charity you find where there is really, really good evidence that it can help people for very little money? Where do you find the charity? And where do you find the evidence?

HOLDEN KARNOFSKY: I could talk about the literal first we ever recommended, but instead, I'll talk about the first where it really started to come together, that we felt we were nailing these criteria. So that was the Against Malaria Foundation. And what they do is they basically distribute insecticide-treated bed nets to help reduce the burden of malaria. Now, insecticide-treated bed nets cost about $5. They often cover two people. Sometimes they cover one person. They last for several years. And they can kill mosquitoes and block mosquitoes so that people don't get malaria.

There's a large number of very rigorous, randomized controlled studies on bed nets. And if you look at the effect size, it looks like they are helping people very cost effectively in the sense of — we estimate that there's a death averted for every few thousand dollars that you can spend distributing bed nets, which is kind of incredible and has been very hard to beat that number at any point during GiveWell's journey.

So, that's the intervention: that's distributing bed nets. And then what we found is we've looked at several organizations that distribute insecticide-treated bed nets. And one of them was really tracking the whole process. They were providing documentation that the bed nets were actually handed out. They were providing shipping information of the bed nets arriving. And over time, they would add more data on going back and surveying people if they were still using the bed nets.

So now you can put together this whole case. We have very strong evidence base. You have an amazing deal, in some sense, where a lot of people get help for each dollar you spend. It's very hard to do that much good helping people in the U.S. with donations. And then you have this organization that is really repeatedly carrying it out and that is providing all the information we need. And so it becomes this very exciting way to spend your money.

EZRA KLEIN: So I want to hold on this idea that giving to anti-malaria charities could save a life for a few thousand dollars, or specifically, according to the GiveWell website, $3,000 to $5,000. How does that compare to more conventional causes that people may be familiar with, like disaster relief or donating to a soup kitchen?

HOLDEN KARNOFSKY: It's about the best we've been able to find, in some sense. We've looked high and low for ways to help people for small amounts of money. Some listeners might be a little confused when they hear $3,000 to $5,000 because they might have heard, well, I saw a charity that can save a life for $0.20.

But at GiveWell, it's always been about rigorously investigating the numbers and subjecting them to all kinds of scrutiny and analysis. And when you really go hard on the analysis and you ask that the number be real, $0.20 to save a life is not something that I think you really have the opportunity to do with your giving. And $3,000 to $5,000, it's just empirically, it's been — I mean, I think that's just incredible compared to most things that are out there.

EZRA KLEIN: What are some of the other numbers you've come to when you've looked at other methods?

HOLDEN KARNOFSKY: A lot of times, we don't put a number on it so much as we just put a bound on it, where we'll say, hey, the evidence here is not very strong. The intervention itself is very expensive. It might be like thousands of dollars just to put a person in a program that might or might not be helping them at all. And so, a lot of times, it's more a matter of saying, this doesn't have a strong enough case, or it's just intuitively too expensive. It's not going to match that other figure.

EZRA KLEIN: So one of the critiques of GiveWell, the GiveWell model, is that, to use the old line about economics, it only looks under the light, right? It looks for its keys, but only where the lamppost is. Because there are a lot of things that could be good in the world, and they don't have a very — it is hard to run a randomized controlled trial of them, because they only happen once, or they are effecting something more diffuse.

So what was the thinking behind demanding this very high level of empirical proof, which, on the one hand, can say, here's where my dollar is going, but on the other hand, you might say, yeah, but the leverage on that dollar is smaller because I'm only trying to save the one life as opposed to influence something, say, 10,000 or a million by averting a war or changing civil service or whatever it might be?

HOLDEN KARNOFSKY: Right, right. I have a bunch of thoughts on this. The original thinking was just very pragmatic. It's just a start-up saying, let's do what we can do. I think sometimes, you do want to start under the light. If there's a big area under the light, well, if your keys happen to be there, you're going to find them a lot more quickly. Maybe you should start there. So I think in some ways, it's not always such a bad thing to do.

Another intuition we had is just that a lot of things that people try to do to help people are just very speculative. They're often based on having certain feelings about the world or feeling like certain things kind of just feel supportive to do. And we thought that things that are really supported by evidence, where you can really drill down and see how much money's going there, might actually be systematically better because they are kind of — they're optimized in a different way.

Or another way you could think of this is the less you know about some intervention, the more you might expect that it's going to just be the average thing you can do. And when you can create a strong evidential case that your money is doing incredible things, that might actually be better than other options.

The final thing I'll say, though, is that I, at least, have branched out a lot since then, so I now run a different organization or co-C.E.O. run a different organization called Open Philanthropy that does not have the same requirement that everything be totally proven based on evidence. And an interesting thing is that we've been looking for things that are better than GiveWell's top charities. And it's been really hard. It's really been surprisingly hard, yeah.

EZRA KLEIN: I'm going to hold you there because we're moving to Open Phil. But I want to do this a little bit more slowly.

HOLDEN KARNOFSKY: Sure.

EZRA KLEIN: You're co-running GiveWell with Elie. It's going well. GiveWell became a big player on the block very quickly. I've given a lot of my money through GiveWell over the past five, 10 years. What makes you decide to split off and start a new organization?

HOLDEN KARNOFSKY: Well, we met Cari Tuna and Dustin Moskovitz. Dustin is one of the co-founders of Facebook and also Asana. And they were trying to give away their fortune in a way that would help the most people possible. And we just felt that a different approach might be called for. When you have a public website making recommendations to anyone and everyone versus working with one family that's giving away a huge amount of money, the second one kind of starts to put new options on the table.

And so, Open Philanthropy still recommends a lot of donations to give those type of charities. And Cari and Dustin still give a lot of money there. But there have been other things that may be higher risk or may be more in the mode of what I call hits-based giving, which is that if you can get the occasional success that is a really huge win, that might make up for a lot of donations that don't quite have the effects they wanted to. And so, that's what caused us to kind of pivot. And we started Open Philanthropy as a project within GiveWell, and it eventually spun out.

EZRA KLEIN: Before you begin Open Philanthropy, you undertake this big study of the history of the philanthropy space.

HOLDEN KARNOFSKY: That's right.

EZRA KLEIN: Tell me a bit about that study and what you learned from it.

HOLDEN KARNOFSKY: When we started working with Cari and Dustin, I wanted to understand a little bit more about how big philanthropies in the past, what they had accomplished. And I didn't know what I'd find. I wouldn't have been surprised if I learned that, actually, they'd never accomplished anything, and we should go back to doing the simplest stuff we can, because it's that hard to help people.

But that's not what I learned. What I learned is that there have been incredibly impressive successes from philanthropy that I think rank up there with the most important events for human welfare in the last 100 years or so. We actually at Open Philanthropy, we've now named about half our conference rooms — each one is named after a philanthropic success story.

EZRA KLEIN: So what are they named?

HOLDEN KARNOFSKY: One of my favorites is called Green Revolution. And that was when the Rockefeller Foundation funded Norman Borlaug and others to research improving crop yields in Mexico. If I'd been around at the time, I doubt I would have said, oh, improving crop yields in Mexico, that's going to be the biggest success that's ever had by philanthropy.

But it may have been because this is generally now credited with kicking off the Green Revolution, where a bunch of countries went from importing to exporting food. And it's credited with saving a billion people from starvation. Norman Borlaug ended up winning the Nobel Peace Prize because it turns out that this was a very scalable improvement in agricultural productivity. So once they had these crops, they could just breed them anywhere. Agricultural productivity took off in a lot of poor countries. And that kick-started all this economic growth.

Another one of my favorites that I'll just throw in is the pill, the common oral contraceptive for birth control. The work was funded by a feminist philanthropist named Katharine McCormick. And at the time, this is the kind of thing that wasn't going to get funded by the government because it was controversial. And in fact, they weren't able to advertise the pill as birth control originally. Instead, the warning label was the advertisement. They had to put on a warning label that it could prevent pregnancy.

So it was an example of philanthropy being ahead of the curve, doing something that was controversial. But they ended up with something that was transformative and was a huge moment for feminism and human welfare because they were willing to be a little controversial

like that.

EZRA KLEIN: One thing I noticed about both of the case studies you used there as examples of huge wins is they're technologies. The Green Revolution are new kinds of crops. The pill is a medication. I don't think most philanthropy is about seeding and staking new technological development. I'm not saying none is — I, in fact, I have a friend who's working on that kind of thing right now — but not most of it. So is that a problem? Is that one of the places where Open Phil begins to diverge a view that you should actually be doing product development through philanthropy?

HOLDEN KARNOFSKY: Technology is not something I would say is neglected by philanthropy. You're probably right that most philanthropy is not technology philanthropy. There is a lot of philanthropy in science, especially today. It's become a very fashionable thing to put money into.

But I would say it does reflect this idea that if you want to do really incredible things that have massive scale and help tons of people, having innovation, having new technologies that can be copied and used freely by anyone, is a great way to get leverage. It's a great way to just have really big effects. And if you don't have any theory in your philanthropy of how you're getting massive leverage and affecting huge numbers of people, then you may be better off with the most straightforward, cost-effective, proven stuff.

EZRA KLEIN: Well, let's talk about how to put that kind of theory into practice. Give me some examples of what you end up funding through hits-based giving.

HOLDEN KARNOFSKY: I mean, one example is just we're the largest funder in the world of farm animal welfare. That's both attempts to develop alternative foods that can reduce meat consumption, but also corporate campaigns to try and put pressure for better treatment of animals. Over the last few years, basically, every major grocer and fast food company in the U.S. has pledged to go cage free. We've been funding a lot of the work that led to that that was already going when we came, but we've tried to speed it up. And we've also taken that work global and been funding a lot of corporate campaigns globally.

And so, while we are a fairly large philanthropy, we're not the largest, but we're the largest funder of that work because that's one of these things where I think we might look back hundreds of years from now and say, that was the greatest moral issue of our time. That was this unacceptable treatment of these creatures we've now decided are kindred creatures that we should care about. But at this time, this is just not an issue that really tends to fire many people up. And so, we're doing something others won't do. And I believe there's been a lot of impact and there's been a lot of difference made because we're willing to go into kind of a weird cause.

Some other examples, I think we were the first major institutional funder of the YIMBY movement. This is the attempt to advocate for less restrictions on building housing to make housing more affordable. And this is, again, just something that was kind of new and weird and has now become a nationwide movement. And that was not how it was when we started funding it.

We funded macroeconomic stabilization policy. So this is a bit of a wonky one, but how does the Federal Reserve prioritize full employment versus controlling inflation? We believe that this is one of the most important things in the whole world for the welfare and bargaining power of the working class. And yet, it's an issue that people often ignore. They think of it as a technocratic issue. Whatever, the experts at the Federal Reserve will decide what to do. They don't see it as an area for philanthropy. And I think we are one of the only philanthropists in that area when we came in and still now. But I think it's tremendously important and we funded a lot of analysis and even advocacy on how to trade off full employment and controlling inflation.

A final example is we've had a biosecurity and pandemic preparedness program since about 2015. And I'm certainly not going to say I've been happy with the preparation response for Covid. But I think it could have been worse. And I think the organizations that have played important roles, by the time we all knew about COVID, it was too late to come in and support them for years and help them be in a solid position and build up a deep bench in a lot of expertise. And it was back when that was kind of an unusual cause for philanthropy to be in that we were supporting all that work.

EZRA KLEIN: Let me ask about a couple of these, though. Macroeconomic stabilization is an interesting one because one way of asking that question is, why do you think you all understand macroeconomic stabilization better than the Fed and others? You're a young organization. You're a young guy. You're not an economist, nor most of the people who work for you. A lot of people worked on this for a long time.

You're coming in and saying there's a huge untapped opportunity here. I would understand for a very explicitly political organization to come in and say, oh, we think the Fed has gotten it wrong. We want more full employment. But what is the difference you have convinced yourself you can make on it?

HOLDEN KARNOFSKY: Well, it's important to understand the general structure of Open Philanthropy is that we consider our expertise in finding causes that are important, neglected and tractable. Tractable means there's something for us to do. And so we try and find the right problems to work on. That's what we consider our comparative advantage. And then when we are doing work, we are hiring and we are funding experts.

And so, this is not about Holden going and learning all about macroeconomic policy and then going and explaining to the Federal Reserve that they've got it wrong. That's not what happened. We funded groups that have their own expertise, that are part of the debate going on. There are experts on both sides. But we funded a particular set of values that says full employment is very important if you kind of value all people equally and you care a lot about how the working class is doing and what their bargaining power is.

And historically, the Federal Reserve has often had a bit of an obsession with controlling inflation that may be very related to their professional incentives. And so we do have a point of view on when there's a debate among experts, which ones are taking the position they're taking, because that's what you would do if you were valuing everyone and trying to help everyone the most, versus which you're taking position for some other reason. So we didn't roll our own macroeconomic policy insights. We funded experts, we funded think tanks. But we do have a point of view on what kind of values should be driving that expertise.

EZRA KLEIN: I think something striking about that list is the sheer diversity of things you all fund. Not only in terms of causes but categories of causes. And this gets to what I think of as one of the most interesting things Open Philanthropy does, which is the way you intentionally divide up your giving portfolio into buckets based on really different ethical, arguably even metaphysical, assumptions. So tell me about worldview diversification.

HOLDEN KARNOFSKY: I need to start with the broader debate that worldview diversification is a part of. At Open Philanthropy, we like to consider very hard-core theoretical arguments, try to pull the insight from them, and then do our compromising after that. And so, there is a case to be made that if you're trying to do something to help people and you're choosing between different things you might spend money on to help people, you need to be able to give a consistent conversion ratio between any two things.

So let's say you might spend money distributing bed nets to fight malaria. You might spend money getting children treated for intestinal parasites. And you might think that the bed nets are twice as valuable as the dewormings. Or you might think they're five times as valuable or half as valuable or ⅕ or 100 times as valuable or 1/100. But there has to be some consistent number for valuing the two.

And there is an argument that if you're not doing it that way, it's kind of a tell that you're being a feel-good donor, that you're making yourself feel good by doing a little bit of everything, instead of focusing your giving on others, on being other-centered, focusing on the impact of your actions on others, which you can get from there to an argument that you should have these consistent ratios.

So with that backdrop in mind, we're sitting here trying to spend money to do as much good as possible. And someone will come to us with an argument that says, hey, there are so many animals being horribly mistreated on factory farms and you can help them so cheaply

that even if you value animals at 1 percent as valuable as humans to help, that implies you should put all your money into helping animals.

On the other hand, if you value them less than that, let's say you value them a millionth as much, you should put none of your money into helping animals and just completely ignore what's going on factory farms, even though a small amount of your budget could be transformative.

So that's a weird state to be in. And then, there's an argument that goes, but even more than that — and this idea is called long-termism — if you can do things that can help all of the future generations, for example, by reducing the odds that humanity goes extinct. Then you're hoping even more people. And that could be some ridiculous comic number that a trillion, trillion, trillion, trillion, trillion lives or something like that. And it leaves you in this really weird conundrum, where you're kind of choosing between being all in on one thing and all in on another thing.

And Open Philanthropy just doesn't want to be the kind of organization that does that, that lands there. And so we divide our giving into different buckets. And each bucket will kind of take a different worldview or will act on a different ethical framework. So there is bucket of money that is kind of deliberately acting as though it takes the farm animal point really seriously, as though it believes what a lot of animal advocates believe, which is that we'll look back someday and say, this was a huge moral error. We should have cared much more about animals than we do. Suffering is suffering. And this whole way we treat this enormous amount of animals on factory farms is an enormously bigger deal than anyone today is acting like it is. And then there'll be another bucket of money that says, animals? That's not what we're doing. We're trying to help humans.

And so you have these two buckets of money that have different philosophies and are following it down different paths. And that just stops us from being the kind of organization that has stuck with one framework, stuck with one kind of activity.

EZRA KLEIN: Before we move on, I want to unpack this a little bit more. So let's focus in on animals for a minute. You alluded to the fact that even if you assign a very low moral worth to animals or to their suffering, 1 percent or 0.1 percent of that of a human, that it ends up adding up to quite a lot. Can you run through that math for me and its implications?

HOLDEN KARNOFSKY: Well, the math would be that — I mentioned before that if you're distributing insecticide treated bed nets, you might avert the death of someone from malaria for a few thousand dollars, which is pretty amazing. And it's going to be very hard to find better than that when you're funding charities that help humans. However, with the farm animal work, for example, the cage free pledges, we kind of estimated that you're getting several chickens out of a cage for their entire lives for every $1 that you spend.

And so this is not an exact equivalence, but if you start to try to put numbers side by side, you do get to this point where you say, yeah, if you value a chicken 1 percent as much as a human, you really are doing a lot more good by funding these corporate campaigns than even by funding the bed nets. And that's better than most things you can do to help humans. Well, then, the question is, OK, but do I value chickens 1 percent as much as humans? 0.1 percent? 0.01 percent? How do you know that?

And one answer is we don't. We have absolutely no idea. The entire question of what is it that we're going to think 100,000 years from now about how we should have been treating chickens in this time, that's just a hard thing to know. I sometimes call this the problem of applied ethics, where I'm sitting here, trying to decide how to spend money or how to spend scarce resources. And if I follow the moral norms of my time, based on history, it looks like a really good chance that future people will look back on me as a moral monster.

But one way of thinking, just to come back to the chickens question, one way of thinking about it is just to say, well, if we have no idea, maybe there's a decent chance that we'll actually decide we had this all wrong, and we should care about chickens just as much as humans. Or maybe we should care about them more because humans have more psychological defense mechanisms for dealing with pain. We may have slower internal clocks. A minute to us might feel like several minutes to a chicken.

So if you have no idea where things are going, then you may want to account for that uncertainty, and you may want to hedge your bets and say, if we have a chance to help absurd numbers of chickens, maybe we will look back and say, actually, that was an incredibly important thing to be doing.

EZRA KLEIN: I want to note something here because I think it's both an important point substantively but also in what you do. So I'm vegan. Except for some lab-grown chicken meat, I've not eaten chicken in 10, 15 years now — quite a long time. And yet, even I sit here, when you're saying, should we value a chicken 1 percent as much as a human, I'm like, ooh, I don't like that.

To your point about what our ethical frameworks of the time do and that possibly an open-field comparative advantage is being willing to consider things that we are taught even to feel a little bit repulsive considering, how do you think about those moments? How do you think about the backlash that can come? How do you think about when maybe the mores of a time have something to tell you within them, that maybe you shouldn't be worrying about chicken when there are this many people starving across the world? How do you think about that set of questions?

HOLDEN KARNOFSKY: I think it's a tough balancing act because on one hand, I believe there are approaches to ethics that do have a decent chance of getting you a more principled answer that's more likely to hold up a long time from now. But at the same time, I agree with

you that even though following the norms of your time is certainly not a safe thing to do and has led to a lot of horrible things in the past, I'm definitely nervous to do things that are too out of line with what the rest of the world is doing and thinking.

And so we compromise. And that comes back to the idea of worldview diversification. So I think if Open Philanthropy were to declare, here's the value on chickens versus humans, and therefore, all the money is going to farm animal welfare, I would not like that. That would make me uncomfortable. And we haven't done that. And on the other hand, let's say you can spend 10 percent of your budget and be the largest funder of farm animal welfare in the world and be completely transformative.

And in that world where we look back, that potential hypothetical future world where we look back and said, gosh, we had this all wrong — we should have really cared about chickens — you were the biggest funder, are you going to leave that opportunity on the table? And that's where worldview diversification comes in, where it says, we should take opportunities to do enormous amounts of good, according to a plausible ethical framework. And that's not the same thing as being a fanatic and saying, I figured it all out. I've done the math. I know what's up. Because that's not something I think.

EZRA KLEIN: I'm struck by that. I really like worldview diversification as a way of thinking about things. And I think it's also relevant as an individual practice. Something I see in my travels around the world, the internet, is people are very intent. Even if they would not say they are 100 percent confident in their worldview, their political ideology, their whatever, they are really interested in making it dominant against all comers. So, just tell me a bit about organizationally, intellectually, the discipline of maintaining a certain level of agnosticism between worldviews whose differences you can't really answer.

HOLDEN KARNOFSKY: So one of my obsessions is applied epistemology, which is like just having good systems for figuring out what your beliefs are in kind of an overwhelming flow of information that is today's world. And I think one of the tools that some people use for it that I find really powerful and I'm going to write about is what I call the Bayesian mind-set, which is this idea that when you're uncertain about something, you can always portray your uncertainty as a number. And you can portray it as a probability.

There's thought experiments. There's tools for doing this. You can say, instead of something is true or false, that it's 30 percent. And you can look back later and you can see if things that you said were 30 percent likely come through 30 percent of the time. I think this is a very powerful framework. And using it can often get you out of the head space of believing that things are true or false and just having degrees of belief in everything and often taking something very seriously, even when you think it probably won't happen, just because it's important enough and it has a high enough probability that it deserves your attention.

And on the other hand, I think this framework sometimes can take people back into a state of fanaticism, where you might say, hey, here's something that would be a really huge deal. And it's at least 1 percent likely. So that means it should be the only thing I think about. It should be my obsession. It's like the examples I was giving before. And that, I think, just lands you in a similarly dogmatic place.

And so, Open Philanthropy is kind of operating two levels of uncertainty. It's often using this Bayesian mind-set. But when the Bayesian mindset brings you to this implication that you'll have to be all in on one thing or another, we'll say no to that, too. And then we'll just go to another level of diversification. And we'll have different buckets with different philosophies on the world.

EZRA KLEIN: I want to pick up on the fanaticism component. And I'm not accusing anybody here of fanaticism. But one of my critiques of the effective altruist world is that it can get very obsessed by that conversion number you were talking about a minute ago. And in particular, I think, it's a culture as it has matured a bit more. There's now an aesthetic, sometimes, of being willing to take the most hardhearted logic experiment seriously and show that you're the real effective altruist because even though it sounds like a kind of terrible thing to do, you ran the math, and it's not.

And the way I'll put this is, Will MacAskill, who's a philosopher and was a founder of the effective altruist movement, used to have this thought experiment where there's a building on fire. And there's a family in one room who could die. And then there's another room — or I think it was, actually, an attached garage or something — that has a bunch of very expensive art in it. What do you save?

And the point of the experiment originally was you should, of course, save the family. And he was making the meta point that many people are donating to museums, instead of to malarial bed nets. I think now, a lot of effective altruists would answer it the other way, because the point is, well, if that art is worth $500,000 and you can turn that $500,000 into x number of malarial bed nets, that saves more than five lives. And so, of course, you need to do that. And I think that gets you into pretty dangerous territory. But I'm curious how you think about those questions.

HOLDEN KARNOFSKY: I do agree that there can be this vibe coming out of when you read stuff in the effective altruist circles that kind of feels like it's doing this. It kind of feels like it's trying to be as weird as possible. It's being completely hard-core, uncompromising, wanting to use one consistent ethical framework wherever the heck it takes you. That's not really something I believe in. It's not something that Open Philanthropy or most of the people that I interact with as effective altruists tend to believe in.

And so, what I believe in doing and what I like to do is to really deeply understand theoretical frameworks that can offer insight, that can open my mind, that I think give me the best shot I'm ever going to have at being ahead of the curve on ethics, at being someone whose decisions look good in hindsight instead of just following the norms of my time, which might look horrible and monstrous in hindsight. But I have limits to everything. Most of the people I know have limits to everything, and I do think that is how effective altruists usually behave in practice and certainly how I think they should.

EZRA KLEIN: What do you think the limit of that actual thought experiment is, of the just convert lives into money? You can save x number of lives for x number of money. And so if you get more money by getting the money as opposed to saving the lives, you should do it.

HOLDEN KARNOFSKY: I think there's a lot of problems with that argument. And I could sort of go into them. So there's things about setting norms. There's things about following rules so that you don't want to be the kind of person who is constantly behaving in strange, unexpected ways and screwing over people around you because you've got this strange mathematical framework that's going on. So I think there's a bunch of things that are wrong with running in and saving the painting.

But I think I also just want to endorse the meta principle of just saying, it's OK to have a limit. It's OK to stop. It's a reflective equilibrium game. So what I try to do is I try to entertain these rigorous philosophical frameworks. And sometimes it leads to me really changing my mind about something by really reflecting on, hey, if I did have to have a number on caring about animals versus caring about humans, what would it be?

And just thinking about that, I've just kind of come around to thinking, I don't know what the number is, but I know that the way animals are treated on factory farms is just inexcusable. And it's just brought my attention to that. So I land on a lot of things that I end up being glad I thought about. And I think it helps widen my thinking, open my mind, make me more able to have unconventional thoughts. But it's also OK to just draw a line. I think it's OK to look at this art thing and say, that's too much. I'm not convinced. I'm not going there. And that's something I do every day.

[MUSIC PLAYING]

EZRA KLEIN: We've been talking a lot here about how to value animals, but the other big worldview here is long-termism, which has to do with valuing future human lives. So tell me more about that worldview.

HOLDEN KARNOFSKY: So the basic idea of that worldview is that if you can do something today that affects all of the future generations, then you have helped a truly mind numbing number of people. We don't know how many people. We don't know how many people will live in the future. But it could be an extremely large number.

Most things that we can do today are not the kind of thing that we have any reason to believe will still matter a billion years from now. But some of them could be. Climate change could be. Climate change is an example of something that could really, in theory, could imaginatively knock humanity off course forever. And causing it to be less likely that this happens could be the kind of thing that matters for every future generation.

And so, long-termism says there are all these people who don't have any voice today in the actions we're taking that affect them. And so, why don't we take the actions that will still matter a billion years from now? Because they'll have affected that many people, and maybe that's the way to do the most good.

EZRA KLEIN: So one of the critiques of long-termism is it quickly gets you into this kind of mathematical moral blackmail, where, well, if you say, because in the future, human beings could spread throughout the galaxies and there could be a trillion of us, over and over again, there could be a trillion of us, so if you give the future human lives 1 percent of the weight of a current human life or 0.1 percent, sort of anything that makes that future more likely to happen is just an astonishingly good investment that outweighs anything you can do for people today. How do you think about that?

HOLDEN KARNOFSKY: I have a few ways of thinking about this. One way comes back to worldview diversification again. So what we aren't trying to do is find the one master framework that is the one thing. What we are trying to do is find things we can do that may turn out to be hugely ahead of the curve that may turn out to be a really big deal, that may turn out to be the best money we've ever spent. I think long-termism definitely checks that box. And so, Open Philanthropy is never going to be an organization that is exclusively long-termist. But we do put a lot of money into it.

And the way you phrased it is one way of phrasing it. But if we also just phrase it another way and we say, why don't we try to focus our actions and our efforts on the things that might still matter a billion years from today, why don't we try to do the things that optimize for having the best future we can, for bequeathing the best thing we can to the future generations, for having the best overall story of humanity that we possibly can, having a healthy society, a society that makes good decisions, a society that's equitable and inclusive?

I don't know. I don't think that sounds nearly so crazy. And I think if you were to walk around all day, this is something I've increasingly been trying to do myself because we've been doing division of labor at Open Philanthropy, and I've been focusing more on long-termism. But if you just walk around all day, just thinking, well, what is it that's the best for helping the world be a good place a billion years from now? I think you end up doing a lot of really super reasonable things and maybe paying a lot of attention to things that should be getting more attention today.

EZRA KLEIN: Well, talk me through how even think of what the long-term in long-termism is because a critique we might have of just the way human beings are right now — I don't think we're typically doing things to make 24 hours from now the best it could possibly be, right? We're pretty far off of an ideal policy.

So then, one version of long-termism is, let's just think about our children's world. And another is, let's think about 150 years from now, which is pretty far in the future. But I'd feel more confident predicting trends out 150 years, knowing I'll get some of them wrong. You then begin talking about a million years, a billion years.

HOLDEN KARNOFSKY: A billion years, yeah.

EZRA KLEIN: What is long-termism to you? Because the further you go out, obviously, the more the uncertainty begins to bite. So how do you define it? And then how do you work within that uncertainty?

HOLDEN KARNOFSKY: From a values perspective, I think that we should be caring about the whole future. But you raise an important point, which is the big obstacle to doing long-termist work is knowledge. And it's very hard to predict the future. It's very hard to identify actions that will matter that long from now. And so the vast majority of ideas we have, we probably will be wrong. And we'll probably be overconfident about what will actually matter a million years, a billion years, whatever.

So this is a huge challenge. And I think it's one of the downsides of being a long-termist and one of the reasons that I haven't put my whole life into it and never will. But that doesn't mean that we're totally helpless. It doesn't mean that we should just throw up our hands and say, let's just optimize for the next one year because that's the same as optimizing for the next billion years. And climate change is an example of that, where if you're always focused on the next year, you might say, let's burn more fossil fuels because that makes us all better off today.

But it's not some radical state of ignorance. It's not some giant unknown question whether climate change is going to make the long run future better or worse and whether it has the potential to make it a lot worse for a very long time. It's actually a real possibility. And so, I think as a long-termist, to do it well, you have to have taste and judgment. And you have to know when you don't know something, which is almost always, and when you might be on to something that actually could matter for that period of time. That's not an easy thing to do. And it's a pretty young idea. So I don't think we're nailing it, but I think someone should be trying it.

EZRA KLEIN: So I know that you started out in a place that's, frankly, more like where I started out or maybe even where I am on these questions, which is a bit more critical of the idea of long-termism, more critical of the appeal of long-termism within the philanthropic circles you run in. But recently, you've put a lot more emphasis in it. In Open Phil's 2018

update on cost prioritization, you wrote, “We’ll probably recommend that a cluster of long-termist buckets collectively receive the largest allocation, at least 50 percent of all available capital.”

Now, I know those numbers are subject to change. But it speaks to the fact that long-termism is an area you decide to place a lot more emphasis on in recent years. So tell me a bit about your trajectory on that. How did you become more persuaded there? And what did you become persuaded of?

HOLDEN KARNOFSKY: Sure. So as co-C.E.O. of Open Philanthropy, my job is to try to get ahead of the curve. My job is to look for ideas that are not only important but also neglected. And so, I’m always looking for what could be the next big thing that could matter for a ton of people that’s not getting enough attention. And I deliberately seek out people and ideas that can introduce me to things that might do that.

And so, through this, I have encountered the idea that this century, the 21st century, could be the most important century of all time for humanity. And a primary way that might come about is the development of A.I. systems that could cause a dramatic acceleration in science and technology, such that if you were to imagine a radical sci-fi future, a technologically advanced utopia, dystopia or anything between or even maybe a world that’s not run by humans at all — that’s run by AIs with their own non-human compatible objectives, which we can get to — a lot of people think that kind of long-run future is possible, but the right kind of A.I. could bring it very soon, could bring it this century.

And once you think that you could be in that kind of century, now the time scales have collapsed. Instead of trying to make predictions about a billion years from now, we’re trying to make predictions about the specific things that could happen in the next few decades that could matter for hundreds of years, thousands of years, billions of years. And so, when I first encountered this idea, I think it all just sounded too wild and too out there. And I really kind of mostly stuck to the work I was doing. But again, it’s my job not to be too dismissive of ideas that could be extremely important and extremely neglected.

So, over the years, I’ve come to take it more seriously. And in particular, Open Philanthropy has had a team really focused for the last several years on taking this thesis about A.I. and the most important century and poking every angle at it, looking for the weak points, trying to figure out whether this is really plausible. And we’ve gotten to the point where I’m not going to say that this is something I know. I’m not going to say this is something that’s going to happen.

But I am going to say it’s a serious possibility that I think deserves a lot more attention than it’s getting. And the most recent kind of change for me is, I’ve been trying to get my thoughts straight. And so, I’ve started my own blog called Cold Takes, where I’m trying to lay out the

case that we're in the most important century as simply as I can. And more generally, have been noticing that I've been taking on more unconventional views, more views that are important to what we're doing, but that are not widely held.

And I think it's important for me to be writing up those views in a clear and simple way in public, not only to help get clear in my own head about what I believe but also — because if I'm wrong, I want it to be easier for other people to encounter what I'm saying and to show me how I'm wrong. So this is a project that I'm on now, is taking these ideas that could be astronomically important that I've started to take quite seriously and continue poking at them by kind of putting them out there.

EZRA KLEIN: Well, let's talk about that project. We'll go deeper into these questions around A.I. in a minute. But first, I want to step back and talk about the big picture view of your— I always feel like you need drums and horns when you say this — Most Important Century series. One of the things you're really arguing there is that the future could be profoundly unlike the past in a way that it's true that 2021 is unlike 1900, but it's a lot like 1900.

It's still human beings running around. A lot of things are recognizable, a lot of the same religions. You had Democrats and Republicans in 1900. There's a lot happening then that was quite similar. But you're arguing here that 2300 could be basically unrecognizable as a world. So give me the basic case for that, the case for why the future might be wildly different than the past.

HOLDEN KARNOFSKY: This is a big thing that has held me back from taking the Most Important Century idea seriously for many years, is that it just — if it's true, it implies we live in this wild, unusual time. And something I've learned is that if you just look at our time in full historical context, there's so many reasons other than A.I. to think that we do live in a wild time. So I think it's helpful to just kind of situate it in context.

First, there's the past. So the universe is more than 10 billion years old. Life on Earth is more than 3 billion years old. The whole thing of a species that's creating its own technology at all, even stone tools. That's millions of years old. So that's millions versus billions. That's the blink of an eye on galactic timescales. And then human history is also just very packed into the recent past.

So I think almost any metric you look at — economic growth, population, major technological milestones — more has happened in the past few hundred years than in the previous several hundred thousand or several million years. And that kind of points to this idea that there is a sort of acceleration that has occurred. Things have moved faster and faster. And if you simply project that acceleration forward and you say the acceleration continues or it's paused right now, but it comes back, in some ways, things going so crazy and the next few decades being more eventful than everything that came before is a continuation of a trend, not a breaking of it.

Then, if you look to the future, I think there's other interesting observations about what a weird time we're in. Today's level of economic growth is a few percent a year. That doesn't feel that crazy to people. But it is not only an incredibly high level by historical standards, it's a level that doesn't look like it can go on forever. So if you try to project out thousands of years of growth of even 2 percent a year, it kind of looks like you've run out of atoms in the galaxy. You just can't do it. And so something has to change.

EZRA KLEIN: Hold there for a minute because I know the math on this, but it's very unintuitive. So, as I remember the calculation you run, 2 percent growth a year for 10,000 years will get you to a quantity that you basically can't run with the materials of the galaxy. But 2 percent doesn't seem that big.

HOLDEN KARNOFSKY: Yeah, I mean, the specific quantity is at some multiple of the number of atoms in the galaxy, and there's very good reasons to think we would not be able to get even close to the edges of the galaxy in that time, because you're just constrained by the speed of light, if nothing else. So, yeah, 2 percent, I mean, it's exponential growth. And if you just plot it out 10,000 years, exponential growth tends to be very unintuitive.

Accelerating growth is even more unintuitive and even more explosive. And that is something we've seen accelerating growth or what's called super exponential growth. We've seen it in the past. And if we see it in the future, yeah, I mean, things go to the moon very quickly, and it's very unintuitive. But I do believe it's something that could happen.

EZRA KLEIN: But why this century? What makes this century so important?

HOLDEN KARNOFSKY: So, three basic points. Point one is that the long-run future could be just radically unfamiliar. It could be a radical utopia, dystopia, anything in between. Point two is that the long-run future could become the near-term future if the right kind of A.I. is developed to accelerate science and technology dramatically. And point three is that that kind of A.I. looks more likely than not this century.

And then the bonus point four is that when you put the three together, a natural reaction is that this implies we're in a very special time. And it sounds too wild to be true. But point four is that if you step out and look at our place in history, it looks like we're in a very weird and wild time for many reasons that have nothing to do with A.I. And so we should be ready for anything.

EZRA KLEIN: I think actually just the idea of acceleration here is unintuitive. I think if you know the basic growth story, it's been fairly steady 2 percent growth globally for some time. People hear a lot about great stagnation. They hear about wage stagnation. It's been very hard for a lot of countries to break out of middle income traps. It doesn't intuitively feel like we are in an era of globally accelerating growth. So, walk me through the accelerationist argument.

HOLDEN KARNOFSKY: Yeah, I think we're not in a period of globally accelerating growth. And I don't think we necessarily will be again. We could just end up with a world that stagnates, where growth slows, like you're saying. The case for accelerationism would be — so the basic idea is, if you look at most standard economic growth models, there's this potential for a feedback loop. This is something that can happen. It's not something I think is happening today.

But you could have a feedback loop where every time you get more resources, more food, whatever, that leads to more people. More people have more ideas. More ideas leads to innovation, and therefore, more resources. So you get resources, people, ideas, resources, people, ideas, and you get a feedback loop. And that causes accelerating growth that can be this very explosive dynamic.

And it looks reasonably likely that this has described periods of economic history. The people refer to the Malthusian dynamic, where people didn't get much richer, even though they were improving productivity, because the additional resources would just go into more people. And so, you see that pattern.

And what happened, what changed a couple hundred years ago or so is called the demographic transition where it stopped being the case that more resources meant more people. And now, of course, when people have a lot of wealth or have a lot of resources, they tend to just be richer. It doesn't cause them to have more children. And so —

EZRA KLEIN: It does the opposite in fact. They have fewer children.

HOLDEN KARNOFSKY: Exactly. And so if that's the dynamic we're in, then, yeah, you would, by default, expect that we will get stagnation. And I think that's a serious possibility that our long-run future looks like economic growth has to slow dramatically. I think there's a lot of reasons to think that is our default. The question is, is there a way to get that same function provided by people, provided by something else, such as A.I., that can be straightforwardly produced by more resources.

So today, we have A.I. systems that are cool. They can beat people at chess. They can transcribe audio. What they're not doing is they're not doing that innovation part of the feedback loop. They're not creating new technologies autonomously. They're not advancing science and technology on their own because there's such limits to what they could do. And the question is, if that changes, if AI is developed that could be as good as humans — doesn't have to be better — at sort of pushing science and technology forward, then you get the feedback loopback back. And then you could get an explosion.

EZRA KLEIN: So there are a number of things I want to question in this or describe more. But I want to first start at what I think will be the natural objection, which is something you didn't spend time on there. In a population-resources-ideas feedback loop, resources is a part of this. So I think it's gotten the worst name recently.

There is a broadly held view — I hold it — we talked about climate change already in the show— that we've had a lot of growth that is using up a lot of resources in a way that, for all the good it's done, has also done a lot of harm. It's put us in a very dangerous climate situation. It has led to a tremendous amount of extinction of other species.

The idea that you would just accelerate growth from there, well, with what resources? Oftentimes, growth does not create resources. It consumes them, right? We're consuming fossil fuels that are finite on this planet, et cetera. How do you think about the resource constraint? Or do you not see it as a constraint? Talk about that piece before we go into what would happen if you blew up the rest of the loop.

HOLDEN KARNOFSKY: First, I just want to be clear that I'm not talking about this possibility as this is exciting, and we got to do it. I'm talking about it as this is something that might happen, for better or worse. And I think it could be very good or very bad. But in terms of resources, so humans right now are the only thing that sort of have ideas, that sort of create new technologies that advance science. And humans have a lot of needs.

And there's a lot of resources for humans that are really hard to replace. There's only one planet that we really know of where it seems pretty doable for humans to exist for a long time. And once you have a different kind of technology fulfilling that part of the feedback loop, that kind of technology does not need all the same resources that humans have. What do you need in order to run more A.I.s? You need more computers.

Well, you need some things for that. You need metal. You need electricity. You need cooling. But there's not that much that you need to run more computers. And actually, all the things you need are very abundant in space. And all the things you need to get to space, I think, can be built with those sort of limited set of resources. So I think that's kind of the whole idea, is that it's a lot easier to build more A.I.s than it is to build more humans. And so this loop could quickly get out of control if that dynamic changes.

EZRA KLEIN: I want to put a pin in there's a lot of metal in space because I think it's actually an important part of your vision, a lot of other people's visions that we should come back to in a second. But in terms of here, this is always a question I have about the AI and locking tremendous economic growth question, which is, oftentimes, the constraint on an idea improving people's lives is material in some way or another.

So, for instance, there is a lot we could do with better analysis leading to better drug innovation. But actually, a huge constraint in drug innovation is you need to run a lot of trials on human beings. You need to actually test things in the real world. It slows everything down. It's a big deal. It's not clear to me, in the first approximation, where A.I. helps on that or how much it actually unlocks.

Or in terms of things people might want, houses take wood. Cars need to be built. It is true that the A.I.s themselves would not need a lot of resources. But in order to get a huge amount of pressure on the growth number, which is measuring things the economy actually produces and people consume, a lot of this would have to be built. It's not going to all be digital t-shirts in the metaverse.

And so, how does A.I. create more resources, as opposed to even potentially disastrously using them up? I mean, should I look at resources as fixed, and more growth will just consume them faster? Or should I look at them as like a pie you can expand? How do you think about that part of the equation?

HOLDEN KARNOFSKY: I think my first answer to the question is that if you assume that you can't translate these A.I. resources into human resources, it doesn't change the fact that we're looking at an enormously consequential development. So if you have this kind of dramatic feedback loop and a dramatic increase in sort of the reach of our world, of our planet's artifacts, if there could be sort of this growing population of A.I.s that is expanding throughout space, that's a very high stakes situation for humans. If it turns out there's no way to convert all of that wealth into wealth that makes humans' lives better, well, that's really bad news.

It doesn't mean that this is a nothing burger, though. It kind of might mean this is really bad news. So when I talk about a giant acceleration in science and technological advancement, that doesn't have to be an acceleration in medicine. But it's an acceleration in some sort of general potency, some sort of resources, some sort of ability to make something happen. And the less compatible that is with humans, the worse. Then it sort of becomes up to us.

I mean, if this happened, could we harness it in a way that it was to humans' benefit? Could we take a very large supply of sort of digital minds or A.I.s having ideas and use that to make the world better? Or is it just going to turn into this sort of runaway train? And I think that's a question for us.

EZRA KLEIN: Well, give me a concrete example. What's an example of a way, if we unlocked this boundary we have on creativity and innovation, right? You have 5 billion A.I. minds coming up with cool ideas all the time. When you think about how that could lead growth to go vertical, what are the kinds of things you're thinking might emerge? What problems that we have might they solve? I mean, it's all fine to talk about growth and GDP and A.I.s, but paint the sci-fi utopia or dystopia for me. Let's get tangible.

HOLDEN KARNOFSKY: Sure, I'll start a little bit more tangible, and then I'll get a little bit more out there. So for the little more tangible, we can talk about energy. Energy is definitely something that if you had a lot of minds and you had a lot of resources, you should be able to get a lot of energy. You should be able to get it very cheap. And you should be able to get

it in a way that isn't necessarily involving any greenhouse gas emissions or anything like that because there's just plenty of energy to be had if you had big enough and good enough solar panels. That's something that I think could absolutely come out of this loop.

Then if we talk about health, I mean, you might be right that health is always going to be bottlenecked by human trials. But if you're able to do enough simulations, if you're able to have enough minds on the problem, you really could come up with a very large number of candidate drugs at once. You might have to wait a few years for the clinical trials. But it's sort of the sky's the limit in terms of how much you could improve health.

And so, that's the answer number one, is health and energy. Well, I mean, that's an awful lot. We could go through other parts of the economy, but dramatic changes in health, lifespan and energy would certainly be a big deal.

Now I'm going to go more to the wilder end of the spectrum and the more speculative end because I think it's important for people to not have their aperture too narrow and not be stuck on things that feel like today, when we could be looking at dramatic changes. So I do want to talk about the more radical end of things. It's very important not to have too narrow an aperture. It's very important not to insist on thinking about things that seem real today when we could be confronting a dramatic acceleration.

So, in the Most Important Century series, I try to just give one concrete imaginable example that won't necessarily be the right one of how technology could get super crazy advanced and give us a huge amount of control over our environment. I call the example digital people. This is very related to the idea of mind uploads.

And it's the idea that you could have sort of people made of software, fully conscious beings. They don't have to be just like us, but they would be something that we decided were properly referred to as our descendants, some sort of digital mind that we would look at and say, I care about that digital mind. That digital mind matters. And identify with them. Maybe I think of them as part of our civilization.

And just from that, just from having either a digital running of a conscious human mind — and we could talk about whether such a thing would be conscious or what that would even mean, happy to get to that if you want — or just something else that we identify with for any reason, that could get extremely radical because any kind of digital being — it could be in a virtual environment, which means that it experiences anything that the runners of the environment want it to experience — that can be unlimited good. That could be unlimited bad. There's no reason that digital people need to have any kind of health problems or any kind of lifespan limits.

And I'm not saying this is necessarily a good thing. There are very scary things about the idea of digital people. You can imagine a world where people are able to lock in certain values to their society forever because of the enormous amount of control you would have

over those virtual environments.

So I think digital people are something where if you just imagine it's possible, you visualize it to be concrete, you draw out the implications, it's very easy to get very quickly to a utopia or a dystopia that just goes all across the galaxy and lasts forever. Now, is that literally what's going to happen? No, it's an example of what could happen if technology advances enough. And there's probably many more I haven't thought of.

EZRA KLEIN: I understand the idea of AI.

HOLDEN KARNOFSKY: Yeah, sure.

EZRA KLEIN: Digital people is a slightly different concept.

HOLDEN KARNOFSKY: Yeah, it is.

EZRA KLEIN: What makes you think that that is something that we could do at a level of fidelity that we should be thinking of them as conscious, having rights, people we could simulate on? You have an interesting idea that one way digital people could really help is by accelerating social science. You could see what happens if you make bunches that meditate two hours a day every day, and others don't. But that you could have the fidelity in them that a simulation of a digital people would be relevant to, I guess, we'd call them real people, physical people. Give me the plausibility argument.

HOLDEN KARNOFSKY: I mean, there's many ways you might imagine having digital beings that we cared about for one reason or another, but the plausibility argument I talk about, just the most concrete, is just, what if we could simulate everything that's going on in a brain? And brains are not these enormous rare artifacts. There's a lot of them.

And there just doesn't seem to be a particularly good reason to imagine that we wouldn't be able to find some way of simulating a brain so that it is performing all the same important operations, all the same information processing, turning the same inputs into the same outputs, especially if you're imagining a dramatic acceleration in the population of A.I.s, which could be doing the science, which could be understanding how to run these simulations, which could be providing the computers that we've talked about. You could have as many as you want.

This is not something where I've decided, oh, yes, this is definitely doable. This is just something that's more like, I don't see why not. I think if we were in this limit of having sort of infinite computing resources, I would expect this to happen by default, but maybe it won't. But it's definitely an example of a way the world could go very wacky very quickly.

EZRA KLEIN: This loops us back to that economic feedback idea because this is something I think is really important in your thought. And it is worth a sitting for a minute, which is that you really identify the core constraint on growth as population. And I don't think that's where

people tend to think about it. They think of resources. They think about ideas.

But in a lot of these different ideas, whether it's A.I. — and it does not have to be sentient; it can just be idea-generating A.I. that never has a feeling of its own, whether it's digital people, simulations of us, whatever it is, you're imagining a world where you just completely blow the top off of the population constraint. So tell me a bit about the population constraint as it operates in your thinking. Why do you see that as what would unlock so much potential for humanity?

HOLDEN KARNOFSKY:. This is largely coming from the different models that economists are using to theoretically model growth. And they all have this kind of point of commonality. Not all of them, but it's just very common. And Open Philanthropy, we did a report called "Report on Explosive Growth" where we kind of very broadly surveyed the economic literature on how people model growth. And there's a lot of ways of doing it.

But this feedback loop is just a very common property. It's a very common theme. And it's true that I think most people don't tend to think of population as this essential input into growth. But I think that comes back to long-term, short-term again. A cool paper that I recently wrote about, there's a paper by Chad Jones that just talks about what you think about long-run growth if you're using what he calls the semi endogenous growth models.

And it does say that over the last 100 years, a lot of our growth has not been about population. It has not been about population growth. It's been about increasing education. It's been about reallocation of resources. But these things, they have limits to how much more of them we can do. In the long-run, it's kind of hard to see how you get continuing high growth or continuing improvement in living standards per capita if you're not having a rising population.

So I think the idea of population is that humans are the only things that produce ideas, innovations, new technologies. You need them to do that. You need more of them if you want more of that. If ideas are getting harder to find over time, then you need even more of them over time. And so, this becomes this very important factor.

[MUSIC PLAYING]

EZRA KLEIN: So I do think taking the dystopic possibilities of all this seriously is important. Something the sci-fi writer Ted Chiang once said to me has stuck in my mind when I asked him about if he thought we'd ever create a sentient A.I. And he said, no, he didn't really think so, but maybe we could. But we shouldn't because far before we ever created sentient A.I. that could kill us all, we would create A.I. that we could make suffer dramatically.

When I hear about this possibility of creating these virtual environments, whether or not it is digital people or just us people, our people, I've always said that the most likely dystopia in the next couple hundred years always feels to me like the Ready Player One dystopia, where

for vast amounts of the population, the actual opportunities in the world have just become so diminished that they are in a Brave New World-like V.R. rig. And things have just become artificial in a thin way as a way of escaping from a grim reality.

When you say you're not sure if this is good or bad, I think that the implication when we talk about growth is that it's all going to be great. What are the things that worry you here?

HOLDEN KARNOFSKY: Things that worry me, number one would be misaligned A.I.s. So I take the possibility seriously. And I know it sounds wacky and like a sci-fi movie, but I take the possibility seriously that things could just spin out of control. You've got A.I.s that can grow the population in a much faster way than humans. And you've got A.I.s just running the world, setting up whatever they're setting up over the galaxy, based on their own sort of random objectives that came in because they weren't carefully designed and are not compatible with humans.

So that seems like probably the thing that I feel best about saying, let's really make sure we don't get that. That just doesn't seem like a human-friendly world in any way, shape, or form. Then, what are other things I worry about? I mean, in general, I don't like the idea of rushing headlong into this world, of having a world where all the people who have inclusive and humanitarian values are focused on other issues.

Everyone working on A.I. is just someone who loves doing fun, sexy stuff and pushing tech forward. Then we kind of rush headlong into this world where it just becomes a race. And it just becomes — you have this dramatic acceleration. And the people who are the least careful and the most eager to win the race end up running the world. And those are not the kind of people that I hope will end up running the world.

So, those are two things I worry about a lot. And a lot of what feels central to me when I think about what do I hope for, if this is the most important century, is, I hope we find a way to have more reflection than either of those implies. There's this term Toby Ord uses in the book, "The Precipice," called the long reflection. And it's the idea of having civilization take its time, in a sense. Use technology to give itself time to improve and to have moral progress and to not build some incredibly large, powerful, galaxy-spanning civilization before it's something that can be done in a fair, inclusive, thoughtful way.

EZRA KLEIN: So, one of the ideas here that you talk about is also that we're in what I've come to think of as a quarter of existential risk. That until the 20th century, humanity hadn't really created technologies that could possibly wipe out all of humanity without a really tremendous amount of effort, right? You really would have needed a lot of people going around, individually killing everybody else.

But then we created nuclear weapons, and all of a sudden, we could. We've talked a little bit about the possibility of synthetic bioweapons. Misaligned A.I. is something people worry about on this score. This is a big point of Toby Ord's book. We are one of the first

generations to really have the power to end human life. And we're not that wise. Our technology has advanced far quicker than our governing institutions, far quicker than our moral intuitions, I would say.

And so, a point you make is that if you get to the other side of this technological era, if you can diversify humanity into other worlds, into other formats of being alive, like computers, maybe there isn't quite as much existential risk anymore. But between here and there, there's a lot. That also strikes me as a reason this is an important period because we may be more protected from wiping out our civilization in a couple hundred years, but we're not now.

HOLDEN KARNOFSKY: Yeah, that all sounds right to me. There is some future state where we've likely become very technologically advanced, where there's not a lot more advances around to kind of shake things up, where there is a great amount of control over the environment, where we're able to be spread throughout the galaxy, have the stability that comes with that, where we might have the stability that comes with not having aging or dying be necessary or just having those take a very long time.

So there's a lot of reasons I think that you might expect just the world to reach a state at some later date where however things are, that's how they're going to stay, at least for very long periods of time. And I think that could be a very good thing or a very bad thing. It just depends on what that state is. And that's what gives me so much vertigo about the idea of the most important century.

It's just, things speed up a lot. And you get to some sort of radical future. And the seed of it, the starting point, seems like it matters an awful lot. And it's this huge opportunity for things to go very well or very poorly. And I don't feel like there's a lot of seriousness about what to do with that situation.

EZRA KLEIN: So should we maybe just not? Should we maybe just not be trying to build A.I., maybe just not be trying to push some of these technologies forward? Should we maybe just not be trying to grow as much as we try to grow?

HOLDEN KARNOFSKY: I think if it were up to me, we would be really slow and deliberate and reflective about it and sort of have some process for making sure that a lot of people were included in a serious reflection on whether this is a good idea before anything really dramatic and irreversible happens. It's not up to me. I think this is coming. I think the economic incentives to create this kind of A.I. are very strong. The idea of having the whole world just not work on it doesn't seem viable.

I also think there are tremendous potential benefits from A.I. and from all technology. Like I've said, I think that the pattern of economic growth seeming like a good thing does look true over the last couple hundred years and probably is true over the coming decades, at least.

So I just don't know that it's realistic to go the way you would kind of go in a happy dream world. And I think instead, we have to take this as something that's coming our way and think about what to do about it.

EZRA KLEIN: Does the entire vision for you actually rely here on this kind of A.I.? Which is to say that if, for some reason, it just turned out that you couldn't create this kind of transformational A.I., that it was never able to learn on enough real world data to create real world innovations, that the computing was too difficult — I don't know — something was found, and it just doesn't — there's a limit we hit and we can't go beyond that on creating computer intelligence, does then the future just look pretty normal to you?

HOLDEN KARNOFSKY: It looks a lot more normal. It would definitely be a big change in my views. It would remain the case that we live in a very strange time. And I feel like we should be ready for wild stuff to happen next. It would remain the case that anything that kind of went into that feedback loop or dramatically accelerated science and technology for any other reason could be enormously high stakes.

There are other technologies I could name that I would consider to be still extremely important, still extremely high stakes. I think the climate change situation would be one of the biggest things looming in sort of this could make a permanent difference kind of zone. So would the risk of pandemics. So I think there'd still be a lot to think about. I think we would still live in an extraordinarily strange time.

But learning that there's no kind of digital mind that could do what humans do to advance science and technology, including a simulation of humans, a digital simulation of humans learning — that's just off the table — that would dramatically move me toward expecting stagnation and thinking about what to do about that.

EZRA KLEIN: So let's live in that world for a minute. Let's not think about digital people, A.I. minds, none of that. Something you tapped earlier is the possibility of easing the resource constraint, because you are mining the incredibly abundant minerals and metals and asteroids and in other parts of the galaxy. Even before you get to settling all over the galaxy, you actually could get a lot of resources out of things that are just flying around space. How important is that as a set of technologies for opening the resource constraint, as opposed to the population one?

HOLDEN KARNOFSKY: That's not something I have really strong views about. I think there's clearly resources that are important for us that we're not going to be able to just get that way. Having a habitable planet and a friendly climate to humans, you can't just take a bunch of metal and somehow make a friendly climate out of that. So I do think eventually, it should be possible to get all our energy from clean sources.

But I think there is a real possibility that if this A.I. thing turns out not to be feasible, we could be looking at a world where we really do stagnate in terms of living standards, and we don't have as much of a growing positive-sum economy anymore. And that becomes a different kind of challenge.

EZRA KLEIN: Would that be a bad thing?

HOLDEN KARNOFSKY: I would think by default, yes. It seems like you've got kind of a wind at your back if there's more to go around every year, more wealth, more just things that make people's lives better. And there is still an enormous amount of poverty today that is just really severe and is quite material. So it seems like there's still quite a lot of work to do in reducing poverty. It's a lot of what Open Philanthropy focuses on. It's what I've spent most of my career on. And I think it's really important.

So, yeah, I think the idea that you stagnate and the idea that you just have what you have, I hope it would happen after there's no more poverty of the kind that we see in lots of the world today. But even if that happens, I think it makes me nervous to have a world where there's no wind at your back of that kind. And it just, we have what we have, and we're all just fighting about who gets the most of it.

EZRA KLEIN: Well, let me take the other side of the argument for it to be provocative.

HOLDEN KARNOFSKY: Yeah, sure.

EZRA KLEIN: So when I talk to you or anyone else who believes in something like the most important century, I hear a lot of, I'm not saying this is good or bad. I hear a lot of, we could invent misaligned A.I. that wipes out all of humanity. I hear a lot of, maybe we'll create digital people or at least some kind of morally important, morally consequential digital life form. And I look around, and I look at how we treat animals. And I think, well, that may not go well at all.

HOLDEN KARNOFSKY: Agree.

EZRA KLEIN: I look at the creation, of course, of weaponry, which you've not talked that much about here, but that could be unbelievably destructive. And there's, of course, an alternative vision, which is, I had Richard Powers, the novelist, on the show a little bit back, who wrote Overstory, just wrote the beautiful book "Bewilderment." And I mean, his view of where humanity should go is not non-technological, but it is into more of a relationship with nature.

I think we called it, in a nutshell, a scientific animism, that the world is getting richer, but we don't need to be that much richer for there to be a lot to go around. We have the capacity to, over the next 30 or 40 or 50 years, power things through solar if we chose to. I mean, the same constraints on our wisdom that keep us from responding to climate change well now, they bite as hard or harder in the kinds of futures you're talking about.

So, maybe the future we should be working towards is not this unbelievably high risk gamble of spreading out throughout the stars and maybe enslaving A.I. or being enslaved by it, but just trying to chill a little bit, share a bit more, protect the climate. Why should that not be the vision?

HOLDEN KARNOFSKY: One, my answer is it doesn't seem to be an option. I mean, we try to focus on things we can do today that will be helpful for the world we're actually heading toward. And if this kind of A.I. being a giant driver of acceleration happens to come out, I mean, I don't think there's much I can do, sitting here today, to just decide that's not the way we're going to go. In the other world, I mean, I think that vision kind of sounds nice at a high level, but there's just an awful lot of horrible grinding poverty in the world today.

And so I don't think that would be a very appealing vision until people have a basic standard of health care and a standard of living and good nutrition, which a lot of people in the world just still don't have enough calories, are still kind of stunted because they didn't get enough to eat as kids. So I think that's very important. There's a lot of work to do there. And that's not something that's going to happen if we don't see more economic growth.

I also think there is — I don't know, but I think the idea of zero economic growth just makes me nervous. I think it should make anyone nervous. Because as long as we have economic growth, there's at least things that are happening in the world that are positive sum. There are projects you can embark on, where you can say, I'm doing this to make everyone's lives better.

And at the point where the only change in resources is zero sum, the only thing that helps one person hurts another person, that's a different dynamic than we've had over the last few hundred years, during which, as I said, it looks like life has been improving. So it's at least something that makes me nervous. It's just, I don't know what the world looks like in that kind of situation. Maybe it turns out wonderful. Maybe we all learn to just have perfect equanimity and live together in harmony and appreciate what we have. But it does make me nervous.

EZRA KLEIN: But I want to pick on this for another minute because I do think it gets to a constraint people will take seriously in a low growth world and then don't take that seriously in the hypergrowth world — say they do, but don't — which is you were just saying that there are many, many people who don't get enough calories around the world. We could change that.

HOLDEN KARNOFSKY: Sure.

EZRA KLEIN: We could.

HOLDEN KARNOFSKY: Agree.

EZRA KLEIN: We could change our energy systems. Many of the things that make the world we live in at its current state way worse for billions of people than it needs to be are solvable, and we don't solve them. I mean, not only do we not solve them, we are not going to get close to vaccinating lower income countries this year. We could. This needed a couple billion bucks and a good global mobilization. And we're not going to do it, even though we could absolutely afford to do it.

And the same shortsightedness, the same lack of compassion, the same inability — actually, I'd say, an expanded inability— as hard as it might be to see somebody in another country as kin, it will be that much harder to see a digital something rather as kin. Hard as it is to see a cow as kin, again, I think the way we will treat non-player characters in this world could be really scary. And so, if we haven't been able to make something closer to a paradise out of this world, I'm not saying why should I believe we will make the hypergrowth world better. I'm more saying that why should I want to give us a chance to make it as bad as we could.

HOLDEN KARNOFSKY: I mean, first off, I would say that we haven't solved the problems that exist today, but we've made a lot of progress on them. And this is where I agree with the news floating around about how the last few hundred years look for humanity. They look good. It looks like there's been a lot of progress on a lot of things. So, there has been progress. And we do get somewhere. I mean, we don't get immediately to where I wish we were as fast as I think we could. But I'm not sure exactly what you're asking here. Are you asking about why the hyper —

EZRA KLEIN: I'm making you — well, I'm trying to just make you question some assumptions.

HOLDEN KARNOFSKY: Which assumptions do you want me to question?

EZRA KLEIN: Simply the assumption that if you want to say you don't know if all this will be good or bad, maybe the possibility that it will be bad makes it such that we should be stopping people from trying to do it. And that if we can even get the world we have right, a world with this kind of dizzying change may be beyond human capacity to manage in any kind of even basically decent way. And so we should fear it as more likely a dystopia than a utopia or even something in the middle.

HOLDEN KARNOFSKY: Yeah, I basically am just in agreement as a maybe — I mean, as this is a possibility. I mean, I'm just in agreement with you. I'm not sure I have an assumption that contradicts anything you said. I think where I am sitting is I'm saying, do I want my role to be trying to stop people from building powerful technologies to address real problems that exist in the world today because of this risk it might accelerate? I mean, I don't want my role to be that, partly because it would be a strange thing to be doing, based on some grand vision for how the long-run future is going to go that I don't think I have. And I do think we have short-term problems that are worth addressing.

But also, I don't know — I don't think that's a path that looks very promising to try and go down. And the three criteria for Open Philanthropy's work are importance, neglectedness and tractability. So importance is how big a deal is an issue, how many people does it affect. Neglectedness is how much attention is it getting. We'd rather work on issues that get less attention. And tractability is what can we do about it. Do we see a path to victory? Are we going to matter?

Open Philanthropy cares about all three of those. So when I think about the possibility of the most important century, I'm looking for paths and things we can do that will make the best of it.

EZRA KLEIN: So let's say you have somebody listening, and you buy into this, or at least, you buy into it as a real possibility. And you want to try to help make the best of it. What is tractable for an individual here? What does this imply for just somebody's life, if they want to try to live on this time scale?

HOLDEN KARNOFSKY: The first answer I want to give is that I wish I had a better answer. I wish I could just tell you, hey, I've not only figured out what's going to happen, which I haven't — I figured out what needs more attention. But I figured out exactly what to do about it. But I haven't done that either. And one of the things I say in my blog post series is, if you're thinking, hey, this could be a billion-dollar company, maybe the right reaction is, yeah, awesome, let's go for it.

And if this could be the most important century, I think my reaction tends to be just like, ooh. I don't know what to do with this. I need to sit down. I need to think about it. I think there's a number of reasons that it's actually very hard to know what to do. And I think we need more attention and more thought and more reflection because most things that I could name as an attempt to make the most important century better could be really good or really bad, depending on your model of what the most important considerations are.

So I can get to that in a second, but I don't want to be totally gloomy about it. I think there are things that look robustly helpful. There are things that look good. One of them is A.I. alignment research. So just, if we could get to a point where we're capable of being confident that we can build advanced A.I. systems that aren't just going to run the world according to whatever objectives they have, and that is a field that exists. That's a field that people are working on. We fund work in it.

Another robustly helpful thing is just trying to find and empower more people who are seeing the situation as it is, taking it seriously, being as thoughtful about it as they can be, and approaching it from the perspective of what's best for humanity, instead of what's best for them narrowly. So I do think there are some activities. And there's some discussion of it in the blog post series.

EZRA KLEIN: And then finally, always our last question, what are three books that have influenced you you'd recommend to the audience?

HOLDEN KARNOFSKY: First book is called "Due Diligence" by David Roodman. He's my co-worker now, although he wasn't at the time he wrote the book. It's about microfinance, which is very small loans to low-income people. And there was this massive debate about how much microfinance helps people, whether it's the best way to help people.

And "Due Diligence" is an attempt to sort through a very overwhelming debate among experts and academics from looking at it from multiple angles, getting the right answer. I think it's a great book to read about how to think. I think it's better for that than a lot of books about how to think, although I generally think writing is a better way to learn critical thinking than reading anything. I think this is one of the best things to read for that.

Book two is called "The Lifeways of Hunter-Gatherers" by Robert Kelly. I want to be upfront. It has a lot of really dry theoretical parts. There's many parts I skipped entirely. There's many parts that I've read maybe a dozen times by now. It's trying to pull together all the evidence we have about what life is like for people who live in a hunter-gatherer lifestyle, which is often assumed to represent what the distant past of humanity looks like, before the Neolithic Revolution 10,000 years or so ago. And it's interesting to see how someone tries to piece together the clues about our very distant past. And like I've said, I've generally been obsessed about thinking about long time frames and getting my head into that space.

Third book is about the future. I'll recommend "The Precipice" by Toby Ord, which we've talked about. It's about the idea of existential risk. It goes through the different things that might happen that could permanently cut off humanity's future. I think it's a really good read. And then I want to throw in two bonus books that haven't been published yet.

EZRA KLEIN: Do it.

HOLDEN KARNOFSKY: So one of them is "What We Owe the Future" by the Scottish philosopher Will MacAskill. It comes out, I think, in September 2022. And it's about the idea of long-termism. It's about what are the things we could do today that could matter to all the future generations. And shouldn't we care about that, when those future generations don't have a voice in the things we're doing today that affect them?

And the final book does not have a title yet. I can't promise it will ever exist. But it's by my wife, Daniela Amodei. And I would expect it in a few years. And it's about the history of people using science and technology to get better gender equality, so things like the pill, things like formula feeding, so that the whole burden of child care doesn't fall on the woman, things like I.V.F. for more options in reproductive timing, and about how in the future, it's going to be necessary to use more science and technology and to have better science and technology to achieve better gender equality.

EZRA KLEIN: Holden Karnofsky, your blog is Cold Takes, which we will put a link to in the show notes. Thank you very much.

HOLDEN KARNOFSKY: Thank you.

[MUSIC PLAYING]

EZRA KLEIN: “The Ezra Klein Show” is a production of New York Times Opinion. It is produced by Jeff Geld, Rogé Karma, and Annie Galvin; fact-checked by Michelle Harris; original music by Isaac Jones; and mixing by Jeff Geld.

[MUSIC PLAYING]