01-16-2014 conversation between Luke Muehlhauser, Eliezer Yudkowsky, and Holden Karnofsky, about existential risk.

Holden: So what's up?

Eliezer: If you've gone on the record at any length about existential risk, I missed it. Most of what I know about your actual opinions on existential risk is from the Effective Altruism Summit.

Luke: And last time I guessed, I was apparently wrong, so I want to be unconfused about your views on x-risk.

Holden: One debate we could have, which I think would be kind of a silly debate to have, would be like GCRs versus x-risk, but I call them GCRs, and I am interested in GCRs.

Eliezer: Okay.

Holden: I think being interested in GCRs is the right answer. Not: be interested in x-risks. I think that the idea of some very sharp distinction is not very important.

Luke: Do you think that's because we have different values, or because we have different conclusions about strategy?

Holden: It's like too much precision and basically… this relates to some broader points. Let's just start with what's on the record. Shallow investigations, half of them are GCRs. So that's one thing. It's what our research agenda been. Then the other thing is blog posts about GiveWell labs...

Luke: From my perspective, your shallow investigations are on the GCRs that knowably don't matter, according to where most of the value is, I think.

Holden: We're basically going to have shallows on all the GCRs, very shortly. Like we're just doing them all. That's like a large percentage of the shallow investigations we've done.

Luke: I'm just pointing out a reason why the difference between us on x-risk versus GCRs matters in terms of what we spend our time on.

Holden: Yeah. So… maybe I'll just start talking…

Eliezer: Before we start talking, just to check, do you want me to try to pass the Ideological Turing Test for your views on x-risk? I don't get the impression that I could, because I like just now discovered that I should have been googling "global catastrophic risk" rather than existential risk to find out what your views are.

Holden: Right. You could try. I think your odds are pretty low, so probably not the best use of time. But if you want to try, I'd be happy to hear it.

Eliezer: Not really. My own impression is that my odds are low.

Holden: So why don't I just start rambling?

Luke: Sounds good.

Holden: Here's some kind of chart of living standards / human power over our environment.

(Holden draws a curve on the white board that is flat for a long time and then spikes up.)

Luke: Sorry, what's on the axes?

Holden: The x-axis is time. The y-axis is ...

Eliezer: Anything! (Laughing)

Holden: Yeah, I don't know. Stuff?

(Laughter.)

Holden: It's some big concept of living standards/human control over the environment.

Luke: Okay.

Holden: Just like, how good we're doing as a species.

Luke: Yup.

"N lives" hypothesis defined

Holden: I'm going to start this conversation just going ahead and presuming that... so there is this claim, I think is very contentious, that if we can colonize the stars, we'll get ... some ridiculous number of human lives, and we should value each of those — causing it counterfactually to exist — as if we were now

saving a life. There is a lot of reasons that's contentious. One, is it's complete guesswork and speculation about how many humans there are going to be. Two, is it's this philosophical thing, about: do you accept the Repugnant Conclusion, do you think that creating life is the same as saving a life ...

Eliezer: To be clear, when you're stating that, were you trying to pass the ideological Turing Test for our own position, or is that your position ...

Holden: Yeah, if something I said there was uncharitable, you should let me know. I think what I just said is like a pretty fair characterization of...

Luke: It does not represent my view.

Holden: Okay.

Luke: For example, in terms of: to what degree is it guesswork that there could be massively more sentient beings.

Holden: I think I was clear that that was my statement. That's guesswork. So let's just be clear, there is this hypothesis, H, that the far future is worth some ridiculous number, I like this one.

(Holden writes 3^^^3 on the white board.)

Eliezer: What? No way can get that number!

Holden: I know, I know, I know. Alright, fine. I'll use n… it was something like 10 to the 50th, but that's almost the same as 3^^^3.

Eliezer: What???

Holden: I'm just kidding. (Laughing).

Eliezer: (Laughing). Well, even if it was just a joke, that's certainly the wrongest thing I've ever heard you say.

Holden: So there is this hypothesis that the far future is worth n lives and this causing this far future to exist is as good as saving n lives. That I meant to state as an accurate characterization of someone else's view.

Eliezer: So I was about to say that it's not my view that causing a life to exist is on equal value of saving the life.

Holden: But it's some reasonable multiplier.

Eliezer: But it's some reasonable multiplier, yes. It's not an order of magnitude worse.

Holden: Right. I'm happy to modify it that way, and still say that I think this is a very questionable hypothesis, but that I'm willing to accept it for the sake of argument for a little bit. So yeah, then my rejoinder, as like a parenthetical, which is not meant to pass any Ideological Turing Test, it's just me saying what I think, is that this is very speculative, that it's guessing at the number of lives we're going to have, and it's also very debatable that you should even be using the framework of applying a multiplier to lives allowed versus lives saved. So I don't know that that's the most productive discussion, it's a philosophy discussion, often philosophy discussions are not the most productive discussions in my view.

So let's just accept this anyway, for a while. The way that I would then characterize this… So let's say that we have, this is history to date, it looks something like that and we need to get here.

(Holden draws a line far above the current peak of the graph.)

Holden: So that means we have a lot more work to do and it means that something that interferes with this trend is really bad. So a GCR is something that can make us go like this, right?

(Holden draws a new line showing the exponential curve slumping back down toward the x axis.)

Holden: So that versus an existential risk should not be considered a huge deal. Besides the fact that it's incredibly speculative to talk about what things are existential versus catastrophic risks, once you get up to the level of unprecedented global damage.

Eliezer: Meta point. Should we sort of like interject where our argument is different than yours, or should we just let you go straight through.

Holden: Let me go straight through for a little bit. I'll pause after not too long.

So actually, this is a good place to stop. There are a couple of things that I get from saying this, one is that it is really important to keep doing this [indicating the upward trend], that's really important and we shouldn't take it for granted necessarily,

although you can argue to me that we should, but I don’t believe we should.

Two, is that this distinction between GCR and x-risk is just not a big deal. Things that cause unprecedented global damage can have knock-on effects that are very hard to predict, for example, war caused by climate change, and it’s kind of nonsense in my opinion to say well, “this thing could kill everything and that thing couldn't,” based on these very speculative beliefs.

But the basic point is that even if you don’t accept what I just said, the [human progress slumping back to the x-axis] dynamic should be about as scary as extinction, because this [upward trend] is very short lived in the scope of human history, and so we shouldn't be confident that something that derails it doesn't involve a risk of us really just falling off the trend and not getting back.

Eliezer: So to be clear, you think that it’s not much worth distinguishing between global catastrophic risk and existential risk because any global catastrophic risk has a probability of turning into what we would term an event which permanently curtails the expansion of intelligent life. So you think that things that cause sufficiently large catastrophic global declines are very likely to turn into never colonizing the galaxies, ever.

Holden: They don’t have to be very likely, because of the numbers we talk about.

Eliezer: From our perspective, if we set the utility of all future galaxies at one, then something with a 90 percent chance of curtailing it is very, in practice, different from something that's a one percent chance of curtailing it by a ratio of 90 to one.

Holden: No, I agree with that math. I see existential risks as really low probabilities in general. I see the things you guys call existential risks as having most of their probability mass concentrated in non-extinction, immediate catastrophic risk anyway. So I think given the level of precision we're working with… I guess you could say to me that AI is a super-special case, and you probably will, but if you take something like bioterrorism, which I know is one that is taken more seriously in the community, I think a bioterror outbreak looks like this, where this is the GCR part and this is the extinction part. Then you have climate change, which looks kind of, well, you have a lot of GCR and then there is the possibility of war and stuff like that, so I just think we aren’t

dealing with huge differences here, given the amount of guesswork we have to do.

Eliezer: So for the transcript, Holden just drew a pie chart diagram both for bioterror and for climate risk. There was this big pie of global catastrophe and a tiny slice of existential risk for both of them.

Holden: Yeah.

Eliezer: I don’t disagree with those diagrams. I certainly don’t disagree with that diagram for global climate change: for global climate change to cause human extinction, it has to cause a war fought with more serious weapons than global temperature rising. Does that sound about right?

Holden: Sorry, say that again.

Eliezer: In order for global warming to cause a human extinction event, something fairly exotic has to happen. It has to cause a war fought with bioweapons, or it has to cause a total collapse of world civilization and then we have to not claw our way up ever. Who is it that's done a bunch of analyses about that possibility?

Luke: Seth Baum wrote a double catastrophe thing. I don't know if that's what you're talking about. Certainly Carl [Shulman] and a couple others, I can’t remember who else, who have talked about likelihood of repopulating the Earth after getting down to 10,000 people.

Eliezer: Yeah, I haven't looked into that in very detail, because it’s not where I think the problem lies.

Holden: Yeah, I don't think it’s repopulating as the issue. It’s getting back to [the upward trend].

Eliezer: The speculation I heard was that we've already extracted all the easy oil, so our civilization can’t climb back up again, and I was extremely skeptical of that.

Holden: Well, I don't know about the oil. I think there’s a lot of things that have to happen. I think we have to think of this [upward curve] as kind of an unusual, miraculous situation. So I think there’s a lot of ways this could get derailed, if we just have an unprecedented global damage.

Eliezer: So why does the sort of upward curve, I'm not going to use the e-word on it, but why is the upward curve supposed to be

unusual? A lot of standard economic theories will sort of give you literally the e-word growth, literally exponential growth, because you can implant two seeds and get back three seeds.

Holden: Yeah, it's just an outside view. We haven't been doing this for most of human history, and we've been doing it for a very short time. So it's arguably this complicated combination of cooperation norms, and of a certain scientific culture and a certain value on innovation and a certain value on peaceful negotiation and a bunch of stuff that we probably just don't understand at all.

The way that things have seemed to work is that there was this very long period of nothing like this happening, and then there was the Industrial Revolution, and then we've seen it happen in country after country, where it kind of just takes off, but no one really knows why. But it's probably copying stuff over from the Industrial Revolution and it's possible there's just some ingredient in here that we don't understand, that we could lose.

Eliezer: Earth underwent a huge period of human economic growth before the Industrial Revolution. The Agricultural Revolution, I think like the usual figure is that it ended up multiplying population by at least a vector of 100 relative to hunter/gatherer times, and I think possibly more than that.

Holden: Yeah, that's multiplying population and not multiplying living standards and power over the environment and likelihood of colonizing the stars. I think that any kind of conversation about possibly getting here only makes sense if you're doing this [upward trend], and I don't think we were doing this on the chart, with the Agricultural Revolution.

Eliezer: Since we're taking audio, we should probably try to throw in at least a few audio cues into there. So: to get to the stars, we have to go a long distance from the Industrial Revolution.

Holden: In terms of empowerment and in terms of human control over the planet, not in terms of number of people on the planet.

Eliezer: So I think that there is legitimate case to be made for the life of a medieval peasant having few or negative QALYs, and I'm not sure that this point is actually relevant to any of the major points we want to make, so maybe we shouldn't go into it. But I think I might sort of agree with it, but nonetheless, there is going to be some quality of life here, whether positive or negative, being

generated by what we would think of as medieval peasants in horrible suffering conditions and that would be multiplied by the total population. I mean, denying aggregative ethics seems like something that would surprise me greatly, if you said it.

Holden: This is not a graph of total utility. This is a graph of living standards/human control over the environment. The easiest way I can say it is that if we went back pre-Industrial Revolution and never got another Industrial Revolution, we wouldn't really be on this trajectory and we wouldn't get [to the stars]. Yeah, there was a big explosion of population after the Agricultural Revolution, but there wasn't, I don't know, in some vaguer sense: our likelihood of reaching the stars didn't really go up, because just all the resources we had just went into supporting more humans and people kind of stayed in this low living standard status.

Eliezer: So based on this argument so far, it's not clear to me whether or not you think that the future has an expected value immensely greater than that of the past to date. Because previously it was sounding like you were denying that, and now you're sort of using arguments that seem like they're affirming it …

Holden: No, no, no, I said I'm assuming it. I said I'm assuming it.

Eliezer: You're assuming that the future is vastly more valuable than the present.

Holden: Uh-hmm.

Eliezer: But you think that the basic trajectory to get there is, to a first approximation, identical with the question, how do we keep economic growth going for present day Earth?

Holden: But I think we're doing great as long as nothing crazy happens, is the way we put it. I think among the global catastrophic risks, I would list stagnation: losing what we have for some unforeseeable reason because we just run out of low hanging fruit, or run out of ways to do it. But I also think that our odds of keeping it going are pretty good, if nothing crazy happens. And global catastrophic risks are something crazy happening.

Eliezer: Have you read the original existential risk paper? Because bangs that immediately wipe out everyone were one out of four of the categories.

Holden: Yeah, I remember that paper. I don't mean immediately wipe out everyone, but I'm talking about GCRs versus...

Eliezer: Whimper is listed as an x-risk because you never get to colonize the galaxy.

Holden: I think in common usage, it’s not considered that way, and when we say GCRs, we're thinking of things that would be a good candidate for causing a whimper.

Eliezer: I really don’t think that the x-risk community would sort of see it that way in natural terms. From our perspective, x-risk equals astronomical waste, equals anything that stops you from colonizing the galaxy.

Holden: Sure, would you include climate change?

Eliezer: I would include that little tiny segment of the climate change pie where we fall down and never climb up again.

Holden: Right, which is the segment where we need geoengineering, which is why geoengineering is an interest of ours.

Eliezer: Geoengineering, I think you could plausibly deploy in any sort of climate change scenario.

Holden: Sure.

Eliezer: It’s just much more important to work on it if climate change is something where we fall down and never climb up.

Holden: That's what I'm saying, yeah. But also geoengineering is more likely to be something that's highly relevant if the effects of climate change are worse than expected. I think if we get the IPCC projection, I don't think there’s going to be a good case for geoengineering and I think probably nobody will make that case.

Eliezer: It seems to me that our uncertainty about the effects of global climate change, like in terms of temperature, seems like our uncertainty there should be much more narrowly concentrated, we should be more confident about the temperature change. Sure, they say but there is this wide range, but stepping back, we're much more confident about how the climate works than we are about what happens if the temperature rises a lot.

Holden: Yeah, I agree. When I say IPCC projection, I mean the whole thing, the impact on living standards and economic growth and all that.

Luke: Right, but the IPCC places very, very small probability on

anything you would call an existential risk, almost none.

Holden: I don't think they place a probability. The IPCC is trying to …

Luke: Sorry: they say words that suggest there's a very small probability, like: Venus-like conditions being basically impossible.

Holden: I don't think I would agree what you just said, but I think it's probably a detail. I think the IPCC mostly just talks about a range that's like a reasonable confidence interval and kind of mostly doesn't talk about things outside that range. So I don't think that it's doing something like claiming that we have less than a one in a thousand chance of something really bad happening. Maybe it says that on some very specific thing, but I think in general, the IPCC projections is, all right, I'm drawing an average and then there is some kind of confidence interval and that's the IPCC report. If we fall in here, we'll probably fine and we're not going to use geoengineering, and geoengineering is interesting if we end up [in an extreme scenario]. And this is a chart of how bad it is, not what the degree temperature change is.

Luke: I agree, they didn't place a probability on certain broad categories of…

Holden: Yeah.

Eliezer: I think I remember hearing about that and being sort of super-unimpressed with the thought that the IPCC thought they could do economic forecasting as well as climate forecasting.

Holden: Sure. I agree, they can't …

Eliezer: Like the Federal Reserve can't do economic forecasting, and they have better resources than the IPCC.

Holden: I don't think the IPCC forecasts are very reliable, and that's why we're interested in geoengineering. If I thought we could take the IPCC forecast to the bank, all across the board, including the economic stuff, I think that I would be not very interested in climate change as a cause, it's just like: it's real harms, it's big harms, but given the amount of attention that's already going into it and given the time horizon we're talking about, I just don't think it would really be on our radar. The interest in geoengineering is because of [tail risk], which I think is pretty likely, because I don't think the IPCC is very reliable.

Eliezer: It sounds to me like our model of how climate change works as a

global catastrophic risk is basically the same. Does that sound right to you as well?

Holden: Yes. I think that's right.

Eliezer: Okay. Do we want to go to bioterror?

Luke: The other place where I think there might be a disagreement is on the likelihood of having another industrial revolution, if we went down to 10,000 people now. But I don't know if that's the most productive debate to have.

Holden: I think it’s fairly likely, but I think if you're making … I think it’s pretty likely that there is nothing that's going to derail us. I think if you're making a list of things that might, even though it’s very low probability, I think that any disruption ought to be on that list. I think that making a huge distinction between things you can imagine killing everyone and things you can’t, is a little bit silly, especially since any huge disruption poses a risk of global war.

Eliezer: I sort of semi-agree. I think that that abstraction qua abstraction is not the key one, if we were distinguishing between bioterror and global warming on that basis, I'm not sure that that would be a sensible distinction to make, qua that distinction.

Holden: Right.

Eliezer: It does sound like you are modeling our trajectory to colonizing the galaxy, as if it’s sort of continuous as a causal mode with stuff that's going on today. Global climate change is a good archetypal global catastrophic risk. So compared to some other sections of EA x-risk, in which it’s thought that the critical causes of our not colonizing the galaxy, or for that matter, the entire trajectory up to colonizing galaxies, is likely to involve some causal modes discontinuous with current causal modes.

Holden: You mean like a singularity?

Eliezer: I mean things that are not happening yet. We don’t use the term ‘singularity’ anymore, it got overloaded. But an intelligence explosion, or self-improving artificial intelligence is something that could come out of left field relative to what's going on today.

Holden: I agree with that statement.

Eliezer: So I'm now drawing a diagram, of which there is a little node at

the bottom and it can go right or it can go left. The right is business as usual, and the left is something weird happens. And if we draw the galaxy at the top, then I think that the people who tend to be very worried about x-risks, have a model where both the events that derail [the current trend], and the events that get us [to the stars] have a pretty high probability of going through a 'something weird happens' node at some point.

Whereas the sort of view that says that x-risk isn't very much worth distinguishing from other global catastrophic risks and isn't necessarily an importantly different area of study in practice from how do you keep the planet going in general says that most of our causal flow to colonizing the galaxy is along business as usual or never has a discontinuous change going from business as usual, and likewise, the things that derail us from colonizing the galaxy, which are fairly improbable because economic growth is pretty steady, would consist of disrupting the same business as usual path. Does that sound about right?

Holden: Well, I guess. I definitely don't think that this [upward curve] just is complete business as usual, I think there's a lot of plateaus and disruptions and jumps and one of them, at some point, will probably be AGI. So I don't know, if I were to draw what I think is going to be the path, I would draw something like…

(Holden draws a series of reverse-L steps from the peak of the curve up to the horizontal line representing colonization of the stars.)

Holden: Then I don't know what this is [Holden points at one of the steps], I don't know what this is [Holden points at another one of the steps], probably one of them is AGI.

Eliezer: So I wasn't so much talking about the speed of the curve, because I think the sort of branch of futurism that cares tremendously about the curve, and whether or not it's a smooth curve, is actually not the branch of futurism we come from.

Holden: No, I know that.

Eliezer: So this could be, could end up being a mathematically smooth curve because once everyone is smart enough, everyone is doing economics and economics says that there is an interest rate, and you invest so as to get the highest interest rate and all the interest rates balance, or something like that.

What I mean is, if you look at current world economic growth, there is a reason it looks more like a curve that obeys the e-word

than [this series of steps]. That's because it’s made up of a lot of different investments. And investments that have very low returns, people don’t invest in. Like the real rate of return…

Holden: Yes, I see what you're saying. But I also think there will be global disruptions along the way. I think there will be … if you look at the history so far, it hasn’t been that much time and it’s generally been one or two really big deals that have been kind of driving everything. It’s possible that you keep having one big deal emerge, just as the old one kind of falls off, and it’s also possible that you just get a more punctuated curve.

Eliezer: Robin Hanson has studied this much more than I have, and he says that what it looks like is the big changes don’t result in the economic level going up, but they result in an increase of the doubling rate. So if you were to draw it on the log chart, it would look something like... I’m drawing a series of lines that were all connected to each other, but had different slopes, and that would be what the log chart would look like.

Holden: Interesting.

Eliezer: Then Robin Hanson tried to extrapolate the chart to talk about when the next great increase in doubling time should occur, but I don’t actually buy that, because it can’t …

Holden: Right, no, I wouldn't either, because like three points on that or something.

Eliezer: He had like five, but still … but when I was talking about this sort of thing, something weird happens, what I meant was that it goes into different sort of things happening, like artificial intelligence happens, nanotechnology happens, things that aren’t like our present world happen, and their impact on the general population might be effectively sudden because they wouldn't be delivered through standard economic... Our current technological innovations get delivered through these sorts of standard channels, like marketing channels, where the price is high because people are making up for a big R&D thing and they don’t have accumulated experience, so the price is high. A few people buy when the price comes down, due to economy of scales. Lots of people buy, it’s gradual, you get a chance to see it coming.

But instead bioterror could just be, up until now, okay, people have gotten sick, but we haven't really done a lot of geopolitics

around people getting sick in huge quantities.

Holden: Yeah, I agree. I think you want a list of all the things that could be highly disruptive and you want to consider them all risks, and you want to consider them all possibilities, I'm not really sure what else there is here.

Luke: I think we might also disagree on what you can figure out, and what you can't, about the future.

Holden: Yeah, I think that's our main disagreement.

Luke: Because I think we make a list and we think we know some things about the items on that list and therefore we can figure out which ones to focus on more.

Holden: Well, no, I would agree with what you just stated, as stated, but I just think that you are more confident than I am. I also believe we can make a list of things that are much more likely to be disruptive than other things, and then we should go and look into them and pay attention to them, but I just think that you guys are much more confident in your view of what are the things. My feeling is this: my feeling is basically we know very little. It's very important to keep this [upward trend] going. That is not something we should neglect or ignore. So generally helping people is good, that's kind of how we've gotten to where we've gotten, is that people have just done things that are good, without visualizing where they're going.

The track record of visualizing where it's all going is really bad. The track record of doing something good is really good. So I think we should do good things, I also think that we should list things that we think are in the far future or just relevant to the far future that are especially important. I think we should look into all of them. Another point worth noting is that my job is different from you guy's job. You guys are working in an organization that's trying to … it's a specialized organization, it's knowledge production. My job is explicitly to be broad. My job is to basically be able to advise a philanthropist and part of what I want to be able to do is to be able to talk about a lot of different options and know about a lot of different things. I don't think it's a good way for me to do my job, to just pick the highest expected value thing and put all my eggs in that basket. But perhaps that would be a good job for many people, just not for someone whose explicit value-add is breadth of knowledge.

So part of it is the role, but I do think that you guys are much more confident. My view is that we should list things we think we know, we should look into doing something about them. At the same time, we should also just do things that are good, because doing things that are good has a better track record of getting us closer to colonizing the stars than doing things that are highly planned out.

Eliezer: So indeed, if I tried to pass your ideological Turing test, I would have said some mixture of “we can’t actually model the weird stuff and people trying to do good is what got us where we are and it will probably take us to the galaxy as well,” that would have been the very thing …

Holden: You just need to water down a little.

Eliezer: Sure, so: “insofar as we're likely to get to the galaxy at all, and it’s highly probable that a lot of the pathway will be people just trying to do good, so just try to do good and get there.”

Holden: Yeah, and it especially will come from people just doing good, as approximate goal and then having kind of civilizational consequences in ways that were hard to foresee, which is I'm particularly interested in opportunities to do good that just feel big. Even if the definition of big is different from opportunity to opportunity, so like a way to help a lot of animals. A way to help a lot of Africans, a way to help a lot of Americans. These are all, in some absolute sense, it seems unlikely that they could be in the same range, but in some market efficiency sense, they're in the same range. This is: whoa, I don’t see something this good every day, most things is good someone else snaps up, let me grab this one, because this is the kind of thing that could be like a steam engine, where it’s like, this thing is cool, I built it. It’s super cool. Then it actually has civilizational consequences.

Eliezer: So in order to get an idea of what you think Earth’s ideal allocation of resources should be, if you were appointed economic czar of the world, you formed the sort of dangerous to think about counterfactual… or maybe a better way of putting it would be: how much money would you need to have personal control over before you started to trying to fund, say, bioterror, nanotech and AI risk type stuff? Not necessarily any current organization, but before you start trying to do something about it?

Holden: I mean, less than Good Ventures has.

Eliezer: Interesting.

Holden: Like, I think we're probably going to do something. But it depends. We want to keep looking into it. Part of this is that I don't have all the information and you guys may have information that I don't, but I think a lot of people in your community don't have information and are following a Pascal's mugging-type argument to a conclusion that they should have just been a lot more critical of, and a lot more interested in investigating the world about. So my answer is: we're still looking at all this stuff, but my view is that no, existential risks are a great opportunity. There's not nearly enough funding there.

Eliezer: Where did you get the Pascal's mugging thing from, though?

Holden: Conversations I've had with people where they say what do you think about MIRI and I kind of say a bunch of stuff about how I don't agree with the strategy, and I'm not impressed with the organization. This is a while ago, so I've modified some of that now, but not all of it. Then they would say, "yeah, but if there's even a chance…" and that was like always the end of the conversation.

Luke: So to be really concrete, I found a comment on Less Wrong from someone saying they donated to MIRI because of Anna Salamon's back-of-the-envelope talk at the Singularity Summit 2009, which she now disowns the robustness of that argument.

Holden: I'm not accusing you guys of …

Eliezer: I think she actually originally said that it was supposed to be a minimum lower bound, not a … "if there's even a chance" argument.

Luke: She said in the talk that she played around with different numbers and all the estimates came within an order of magnitude, which she now thinks is not robust.

Holden: People constantly, and I'm not saying you guys, but people in the community constantly reason from uncertainty to minimum lower bound, which to me is absurd. I think I know why it's absurd, but even if I'm wrong [about why], it's still absurd. I have no idea if this is going to happen, therefore, I have to model at least a one in a thousand chance.

Eliezer: I've like tried to explain to people what's wrong with that, as a matter of epistemology.

Holden: I would love to know what's wrong with that. Can you write a blog post about it? I mean, I have my theory, but… I was going to write …

Eliezer: I'm trying to think if I already have a blog posted about it.

Holden: Pascal’s Muggle was sort of…

Eliezer: It wasn't exactly the same issues. So basically what they're trying to do is… it’s “we are uncertain about whether we have model error and whether the large Hadron Collider blows up the world, therefore we have to put this lower bound on it.” I just had a really long conversation about this at Oxford, but we didn't record it, and I don't think I have it in a summarized blog post.

Holden: I'd be really happy if you wrote it up and then maybe I'd cancel my post on this.

(Laughter.)

Eliezer: So noted, I do have a pile of stuff to write, obviously, and so do you.

Holden: So anyway, that was an aside. I think you guys are more in the camp of thinking you understand the issues really well and not only understanding what the issues are, but who is working on what and believing that the neglectedness of x-risk is a large part of your interest in x-risk, I think. I think there are a lot of people who reason so quickly to believing x-risk is paramount that I don’t believe they've gone out and looked at the world and seen what is neglected and what isn't neglected. I think they're instead doing a version of Pascal’s mugging. But I'm happy to engage with you guys and just say that I don't know everything about what's neglected and what isn't, I think existential risk looks pretty neglected, preliminarily, but I want to look at more things before I really decide how neglected it is and what else might be more neglected.

Do you agree with me that the neglectedness of x-risk is a major piece of why you think it’s a good thing to work on?

Luke: I think it is for me.

Eliezer: I think I would like specialize that to say that there are particular large x-risks that look very neglected, which means you get a big marginal leverage from acting on them. But even that wouldn't really honestly actually carry the argument.

Holden: But if you read “Astronomical Waste,” it concludes that x-risk is the thing to work on, without discussing whether it’s neglected, and I think that's the chain of reasoning most people are following. I think that is screwed up.

Eliezer: Yeah, that can’t possibly be right. Or a sane Earth has some kind of allocation across all philanthropies. And insofar as things drop below their allocations, you'll get benefit from putting stuff into them, and if they go above their allocations, you're betting off putting your money somewhere else. There exists some amount of investment we can make in x-risk, such that our next investment should be in Against Malaria Foundation or something. Although that actually that still isn't right, because that's a better argument now because GiveWell actually did say Against Malaria Foundation is temporarily overinvested, let’s see what they can do what with their existing inflow.

Luke: Though not necessarily relative to the current allocation in the world!

Holden: Yeah, absolutely.

Eliezer: But there also exists some amount of money GiveWell could get, such that they would start giving Against Malaria Foundation money again, just because …

Holden: Yeah, there does, yeah.

Eliezer: I agree with you that there is a landscape, and… for that matter, MIRI just did a fundraiser that went well relative to CFAR’s fundraiser, and I posted to Facebook saying the next marginal dollar is now more valuable at CFAR.

Holden: Cool.

Eliezer: In other words, some amount of money flowed into MIRI, such that… not that AI has stopped being important, but such that CFAR’s next marginal… like CFAR is what I think of as the cover-all-the-bases thing. Sort of like a create community that can respond to these kinds of things as they happen in the future.

Luke: I think of GiveWell significantly that way as well. People who are learning to think critically about what's good to do in the world, and cause neutral, single magisterium, try to think about everything the same way, etc., those things are going to be really useful 30 years from now, when we need to pivot the

Earth.

Eliezer: Yeah, and I think the most likely case for “the history of the galaxy is written and GiveWell was important somehow,” or me, obviously all the utility comes from there, at least in my book, and I'm not sure where we should go with that.

Luke: If AMF was important to the galaxy, it was because it helped GiveWell.

Eliezer: Yeah, exactly.

(Laughter.)

Luke: That's kind of my view.

Eliezer: Yeah, that's frankly my view as well.

Holden: That's a part of my view. Just so you know, when we started GiveWell, that was … it’s not like we reasoned it out this explicitly or had this much of a plan, but we said: what do we want to do? Analyze where to give, and have people listen. What's the right thing to start analyzing? The thing where there is all the data and we can do something impressive [and useful] without having to … I mean, we couldn't do the work we're doing now back then. It wouldn't have been possible and I can’t [easily/quickly] explain why, but that was a big part of the reasoning. I wouldn't go as far as you guys…

Eliezer: I am quite sure that you think that a lot less than we do, because you drew a path to colonizing galaxy that has AMF on it.

Holden: Sure. I mean, I don't know. I think you also have to keep in mind that I only conceded this [“allowing x lives is worth some reasonable multiplier of saving x lives” hypothesis”] hypothetically. I guess my view is the way that I handle Knightian uncertainty is that I just kind of don’t put too much weight… or rather, I try to make sure that I'm covering other bases. I was talking with [person] about this and I might mangle this, but it’s more of an interest in robustness and robust optimization, than an interest in optimization given my current best guess at each parameter. I think that's actually rational and that leads to optimization over time, because of the way it interacts with learning over time.

I think that for GiveWell to just put it all into AI risk right now, even if that's where …

Eliezer: *I* don't even think that's the best thing to do.

Holden: Yeah. Okay. I think some disagreements we have, which I think are like not enormous disagreements, I think they mostly have to do with how confident we can be. I think we agree that there are many things that are important, we agree that being neglected is part of what makes a cause good. If there are other causes that are really important and really neglected, those are good, too. We agree that everything that is good has some value, but I think the things that are good have more value relative to the things that seem to fit into the long-term plan and that has a lot to do with my feeling about how confident we can be about the long-term plan.

Eliezer: My reasoning for CFAR sounds a lot like this. Why, to some extent, I sort in practice, divide my efforts between MIRI and CFAR, is sort of like this, except that no matter what happens, I expect the causal pathways to galactic colonization to go down the "something weird happens and other weird things potentially prevent you from doing it" path.

I think that human colonization of the galaxy has probability nearly zero.

Holden: Right, you think it would be something human-like.

Eliezer: I'm hoping that they're having fun and that they have this big, complicated civilization and that there was sort of a continuous inheritance from human values, because I think fun is at present, a concept that exists among humans and maybe to some lesser extent, other mammals, but not in the rocks. So you don't get it for free, you don't want a scenario where the galaxies end up being turned into paperclips or something. But: "humane" life might be a better term.

Holden: Sure, sure.

Eliezer: I think that along the way there you get weird stuff happening and weird emergencies. So CFAR can be thought of as a sort of generalized weird emergencies handler.

Holden: There's a lot of generalized weird emergencies handlers.

Luke: Yeah, you can improve decision-making processes in the world in general, by getting prediction market standard or something.

Holden: Also just by making people wealthier and happier.

Eliezer: Prediction markers have a bit of trouble with x-risks for obvious reasons, like the market can’t pay off in most of the interesting scenarios.

Holden: I think you can make humanity smarter by making it wealthier and happier. It certainly seems to be what's happened so far.

Eliezer: Yeah, and intelligence enhancement?

Holden: Yeah, well, that, too. But that's further off and that's more specific and that's more speculative. I think the world really does get smarter, as an ecosystem. I don’t mean the average IQ. I think the ecosystem gets smarter. If you believe that MIRI is so important, I think the existence of MIRI is a testament to this, because I think the existence of MIRI is made possible by a lot of this wealth and economic development, certainly it’s true for GiveWell. If you take my egg and run it back 20 years, my odds of being able to do anything like this are just so much lower.

Eliezer: CFAR, from my perspective, it’s sort of like: generalize those kind of skills required to handle a weird emergency like MIRI and have them around for whatever other weird stuff happens.

Holden: I think the world ecosystem has been getting better in handling weird emergencies like that. I think that part of that, if you want to put a lot of weight on your CFARs, then I think that's evidence, and if you don’t want to put a lot of weight, then I think there’s other evidence. There is more nonprofits that deal with random stuff, because we have more money.

Eliezer: I'm not sure if I'd rate our ability to handle weird emergencies as having increased. Nuclear weapons are the sort of classic weird emergency that actually did get handled by this lone genius figure who saw it coming and tried to mobilize efforts and so on, I'm talking about Leó Szilárd. So there was a letter to President Roosevelt, which Einstein wrote, except Einstein didn't write it. It was ghost-written by someone who did it because Leó Szilárd told them to, and then Einstein sent it off. There is this sort of famous story about the conversation where Leó Szilárd explains to Einstein about the critical fission chain reaction and Einstein sort of goes “I never thought of that.” Then came the Manhattan Project, which was this big mobilization of government effort to handle it.

So my impression is that if something like that happened again, modern day Einstein’s letter does not get read by Obama. My

impression is that we've somehow gotten worse at this.

Holden: I don’t agree with that.

Luke: Eliezer, why do you think that?

Holden: You're also pointing to a very specific pathway. I'm also thinking about all the institutions that exist to deal with random stuff these days. And all the people who have the intellectual freedom, the financial freedom, to think about this stuff, and not just this stuff, other stuff that we aren’t thinking about, that can turn out to be more important.

Eliezer: We don’t seem to be doing very well with sort of demobilizing the nuclear weapons of the former Soviet Republics, for example.

Holden: We're also talking about random response to random stuff. I think we just have a greater degree of society’s ability to notice random stuff and to think about random stuff.

Eliezer: That's totally what I would expect on priors, I'm just wondering if we can actually see evidence that it’s true. On priors, I agree that that's totally expected.

Holden: Well, it also could be false for some other reason related to some particular dysfunction of the last few decades of development or something. But I think all else equal, more health and wealth and peace, we ought to be expected…

Eliezer: We could somehow be worse than it was in the 1940s and yet, still, increasing development could all else equal improve our capacity to handle weird stuff. I think I'd agree with that. I think that I would like also sort of agree that all else being equal, as society becomes wealthier, there are more nonprofits, there is like more room to handle weird stuff.

Holden: Yeah, it’s also true that as we solve more problems, people go down the list, so I think if it hadn’t been for all the health problems in Africa, Bill Gates might be working on [GCRs], or he might be working on something else with global civilizational consequences. So when I'm sitting here not knowing what to do and not feeling very educated in the various speculative areas, but knowing that I can save some lives, that's another reason there is something to that.

But it’s certainly like: the case for donating to AMF, aside from the way in which it helps GiveWell, is definitely in a world in

which I feel very not very powerful and not very important [relative to the world Eliezer and Luke envision]. I feel like, you know, I'm going to do [a relatively small amount of] good and that's what I'm trying to do. So in some sense, when you say, AMF isn't like a player in the story or something, I think that's completely fair, but also by trying to take a lot of donors who are trying to do this much [a small amount] and trying to help them, we've hopefully gotten ourselves in a position to also be a player in the story, if in fact the concept of a player in the story ends up making sense. If it doesn't and this [small amount of good] turns out to be really good, we'll at least have done that.

Eliezer: The sort of obvious thing that I might expect Holden to believe, but I'm not sure that that actually passes your ideological turing test is that collectively, fixing this stuff collectively, is like a bigger player than collectively the people who go off and try to fix weird things that they think that the fate of the future will hinge on.

Holden: I just think it's possible that what you just said is true, and possible that it isn't. If I'm sitting here, knowing very little about anything, and I want to do a little bit of good, I think doing a little bit of good is better than taking a wild guess on something that I feel very ill-informed about. On the other hand, our ideal at GiveWell is to really be playing both sides.

Eliezer: What do you think that I, Luke, MIRI would say. What's your ideological turing test version of our case for x-risk?

Holden: I could think about this harder. My immediate reaction, which I think is very abstract and vague, but probably passes the ideological Turing test, is that you've spent an enormous amount of time thinking about this stuff and have addressed the various objections, like "far future predictions don't have a good track record" in some ways that Holden hasn't fully seen, and that you feel that you've accumulated the degree of evidence and understanding needed to overcome the basic question of "how well can you predict the future" and so your expected value is higher over here [on x-risk]. I don't think anything I just said is unreasonable, which is why I think it probably passes the test, it's pretty vague, though.

Eliezer: I'm not sure I agree with that. My version of this: I think the way the Holden route actually plays out in practice, when I visualize trying to do that, is that you fix malaria in Africa, you may be able to even actually fix that. You move onto the next thing. You

can even sort of fix governance in Africa, global quality of living goes up, there are more nonprofits. Some of them are weird nonprofits, some of those are counterproductive, others are productive.

The great story of the next decades may not even be so much having a big break in the great stagnation, as just like the rest of Earth coming up to First World living standards. So that the entire planet is effectively this like much larger place that you'd think would be able handle weird things and one night an AI undergoes an intelligence explosion and the next day the Earth has been transformed into, not exactly paperclips, but something that en masse we don't think has very much value.

Luke: So far, that's compatible with what's Holden just said.

Holden: Yeah.

Luke: I think what Holden just said is a pretty fair characterization of what I think.

Holden: I think the first thing reduces the probability of dangerous intelligence explosion.

Eliezer: I don't actually see a very strong connection there. We're pretty wealthy now and we seem to be falling down on the job. I'm not sure that we'd stop falling down on the job if we got somewhat wealthier.

Holden: Presuming for the moment that MIRI is a huge step forward, MIRI is made possible by a lot of the wealth and comfort that we have, and I mean, this is just like an incredibly hackneyed argument, but there could be 10 more people who have the strengths you have, who are basically not able to make the most of them, or not willing to make the most of them. Sometimes people get attached to random problems. If Africa didn't have this problem… [Bill Gates might be doing something different today]. I don't know, maybe he'd be doing exactly what you're doing right now, if it weren't for that.

Eliezer: It seems exceedingly unlikely for other reasons. Due to the great stagnation, the First World part of the world that does fun things like MIRI, has not undergone very much economic growth over the timespan in which Earth has gone from entirely one hundred percent neglecting intelligence explosion to devoting *some* resources to it.

Holden: There’s been plenty of economic growth, what are you talking about?

Eliezer: The First World, like the United States, median income stagnation.

Holden: Why are you talking about medians?

Eliezer: Yeah, I suppose that's fair.

Luke: I think just the last 30 years of Silicon Valley, we now have 30 billionaires who believe that you can think about the future. This is way better for MIRI’s existing than …

Holden: Again, just like me, I just wouldn't be … look at my crazy career trajectory. For one thing, I went to Harvard, because their admissions are kind of based on grades and test scores and stuff like that now. I think they used to be based more on other things.

Luke: There was a hilarious anecdote, in this one book, about the president of Princeton saying that the first quality they look for in graduates in 1960 was extraversion.

(Laughter.)

Holden: Yeah, exactly. So first, I was a middle class kid in Chicago and ended up at Harvard. That was a big jump in my ability to do things with my life, that really just came from that admission process working a certain way. Then after that, I went to a hedge fund for a few years, and I was like, “okay, I am making plenty of money, I don’t have a lot of career risks that I need to seriously worry about. I have all these tools for managing my time. I have all these side projects going on…” GiveWell was not the first side project I did.

The fact that I had a bunch of them, mattered. A lot of things got me to be able to do GiveWell and to be willing to do it. Another really random observation, is that [there were other fields I might have preferred if not for the existence of other people whose performance convinced me that I wouldn't be good enough at those fields.]

But if I'd gone into one of those other fields, would I have discovered effective altruism? Would I have thought about it? I don't think these things are so [determined]. So generally people have a lot, especially people that go to Harvard, have a lot more options today, more than they used to, and a lot better ability to

manage their time to think about random stuff and learn about random stuff.

Luke: I think, broadly, Eliezer and I probably agree with “more good stuff happens when you are wealthier,” it comes back, for us, to this argument that depends on belief of our knowledge that you probably don’t accept, about how when you grow the economy, we’re worried that this accelerates parallelizable work like AGI more than highly serial work like Friendly AI.

Holden: Right.

Eliezer: More generally, from our perspective, since there isn't a presently safe looking well-going path to galactic colonization, things like peace and global economic growth gives you more good stuff and more bad stuff. It’s not clear what the scaling factor is. I do feel a bit unsure and non-virtuous for even thinking about saying that global economic growth is a bad thing and so on.

Holden: Yeah, I think I feel more sure than you that the net of global economic growth in expectation is positive.

Eliezer: I would well expect that.

Holden: And I feel less confident than you in the particular pathways you see. The way that I see it, it’s like there is this giant wave and the wave is very good and we need to push the wave. You guys are worried about the wave, because there’s this one fish that's getting pushed by the wave. You may be right, that that fish is so important, but I don’t have that confidence.

Luke: What about the general Moore’s Law of Mad Science, though? So: easier to destroy value structures than to preserve them, and every year that we get smarter in controlling the world, the IQ required to destroy the world drops.

Holden: This is something that I actually did lay out in the blog post, and this is called something like … I don't know, it had global catastrophic risk in the name and I don't think there are a lot of blog posts written that have that. I think this might be a better argument to take off line. I actually think that basically if you're worried about generic mad science stuff, that faster development is safer than slower, even if no development is safer than some.

Eliezer: How would that play out in bioterror?

Holden: Basically, if you assume that faster development includes both development of dangerous things and development of danger-reducing things, and you believe that the danger-reducing things are kind of statistically favored, there are more people working on them: it's basically a faster progression leaves less time for the bad guys to get lucky. This is something that might be worth taking offline. I think if there is a specific risk, that you have a specific plan to execute, then my analysis doesn't work and I think that's where we differ. You guys think there is this one risk and you have a plan and no one else is working on it, and so you want everyone else to slow down while you move forward. And I think that makes sense on its own terms.

Eliezer: It's not like we have to ask people to do that. It's more like, we don't think we can do better by speeding things up.

Holden: Sure, sure, sure. That makes sense when you have a plan. But when you're just worried about generic bad stuff happening, that might also be offset by generic good stuff, you want faster, assuming the good stuff is statistically favored.

Eliezer: I'm not sure I agree with that as a fully generic analysis. Would Earth have been better off if nuclear weapons had been developed faster?

Holden: But that's a weird way to put it. Earth would have been better off if the whole zeitgeist of growth and development had happened faster, I think.

Eliezer: I think in particular, okay, so nuclear weapons are a very special case, first because even by the standards of x-risk, there turned out to be nothing you could do about them, besides develop deterrence shields. Second, because they arrived just as World War II is ending and things would have magically been much worse as a special case if they'd been developed more during World War II.

Holden: Yeah, I think that's true. I think this analysis is still right for the future.

Eliezer: But as near as I can tell, resources don't get invested in prevention technologies until something has gone wrong at a sufficiently large-scale. Then there's this research lead time. So it seems to me that if biotech moved more slowly, in general, both the good parts and the bad parts, then there would be a better

chance that the first major “10 people in a basement home brewed a virus that gets out of control” incident is relatively small the first time it happens, and that produces a great public outcry… Part of it is just the sort of biotech is broken for weird reasons, but that might be generic to other causes, due to great stagnation type stuff.

So, as money flows into me, I think AI research saturates relatively quickly and maybe 10 million dollars per year or something like that, CFAR saturates more slowly than that, but possibly even before either of those has been saturated, I start worrying about the sheer brokenness of biotech and the lack of attempts to develop rapid response capabilities, to engineered viruses and bacteria.

Holden: But this is something that we're talking about. Nothing has happened to raise the public outcry. We're trying to build systems to block this stuff.

Eliezer: What I'm saying is that as it goes faster, as technology develops faster, the first big incident I think gets worse, and there is sort of less time to respond before the next big incident.

Holden: That's interesting, I'll think about that.

Eliezer: I think this is a fully generic thing across the catastrophic risks. I don't think it’s specific to biotech or to AI.

Luke: I just think that it’s a lot easier to build a nuclear bomb than to build a nuclear bomb shield, it’s easier for Craig Venter to build a synthetic virus than it is for people to build the technology required to prevent it spreading, etc.

Eliezer: I do think it’s legitimate to say that nuclear weapons were something of a special case. I think if you look at biotech or AI, then the investment ratio needed to keep parity is very different from the investment ratio needed to keep parity with nuclear weapons.

Luke: Sure.

Holden: You're also talking narrowly about tech danger, tech response. So a couple of rejoinders, one is that a more peaceful world handles this stuff better and a more secure world and a richer world all handle this stuff better. Another rejoinder is that one of the things that inspired me to think this way is the more generic safety improvers, which would be things like, for one, [causing]

humans [to become] better and smarter. For another, just like generic security measures, like increasing interest in security; and for another thing, just big things, good AGI and colonization of the stars. Another difference between us is I just think AGI is overwhelmingly likely to end up being a good thing [if it happens] and I'm sure you guys don't believe that.

Eliezer: Do we think a necessary and sufficient cause of our disagreement is *just* our visualization of how AI plays out?

Holden: I think it's possible.

Eliezer: If your visualization of how AI worked magically, instantly switched to Eliezer Yudkowsky's visualization of how AI worked. I mean, Eliezer Yudkowsky, given sudden magical control of GiveWell, does not just GiveWell to be all about x-risk. Eliezer puts it on the link like three steps deep, and just sort of tries to increase the degree to which incoming effective altruists are funneled toward...

Luke: I wonder if the difference even can be characterized as the difference between Holden's view of how AI works, and how Bostrom's view of AI works, which is even less narrow than yours.

Eliezer: I'm not sure Bostrom and I disagree all that much. Bostrom just says it in a much nicer fashion.

Luke: He has a much broader view about what's useful to do now about it, for example.

Eliezer: Maybe. I'm not sure how much of that is … I'd have to talk with Nick to be sure we actually had disagreements going on there.

Holden: I think it's pretty possible, and I just want to contrast what you guys think with the normal tenor of the arguments I have over x-risk, which… I just talk to a lot of people who are just like, look, x-risk is clearly the most important thing. Why do you think that? Well, have you read "Astronomical Waste?" Well, that's a little bit absurd. You have an essay that doesn't address whether something is neglected, concludes what's most important, and we're not even talking about AI and path to AI and why AI, it's just x-risk, [which people interpret to mean things like] asteroids, come on.

Eliezer: I endorse your objection. We can maybe issue some kind of joint statement, if you want, to inform people.

Holden: Yeah, perhaps. I was going to write something about this, so maybe I'll run it by you guys. To the extent that I'm known as Mr. X-Risk Troll, or whatever, it's because those are the arguments I'm always having. When I think about you guys, I think that you and I do not see eye to eye on AI, and that goes back to that conversation we had last time, and that may be a lot of the explanation. At the same time, it's certainly on the table for us to put some resources into this.

Eliezer: Although I do want to say something aloud along the lines of: be fair to our people, it is very rare, in general, that people have the concept of diminishing marginal utility as one of their fundamentals of effective altruism, and I suspect a lot of people are sort of like moved by "I have so much, I should give to the deworming charity," more than the deworming charity is under-invested in for our total planetary distribution of philanthropy. I'm not sure this is particular to x-risk…

Holden: I think GiveWell fans, the really big ones, are really into marginal dollars.

Eliezer: I think you spend a lot more of your day talking to GiveWell's smartest fans than you spend talking to MIRI's smartest fans.

Holden: Yeah, that's true.

I mean, I think [Person 1], for example, I remember pretty clearly, years ago, and so she may be different now, but that was definitely the argument I had with her, and I was just like, oh, what do I do with this? She's pretty smart, she's up there. I think this was the tenor of the first conversation I had with [Person 2] on the topic, that was many years ago, and he talks about it very differently now. Probably I would say exactly what I just said about [Person 2], I will also say about [Person 3], so I'm not talking about random…

Eliezer: I have had this argument with [Person 3], where I'm like, [Person 3], stop telling people about these low probability arguments, please don't do that.

Holden: And Nick Bostrom himself wrote "Astronomical Waste," so the people I'm picturing are like [Person 2] 2007, [Person 3] 2008, [Person 1], like 2008 or 2009. Nick Bostrom, whenever he wrote that essay.

Luke: Certainly aren't random people.

Holden: Yeah.

Eliezer: I agree with that. Though I'm not sure if Nick Bostrom made that argument.

Luke: Even if you're writing a philosophy paper, where it's not correct to write about empirics, you can still throw in a few lines about how these other questions matter.

Holden: Yeah, yeah, definitely. He was also like… the paper just had no caveats. It just wasn't … it was like, "I have proven this."

Eliezer: I'd have to reread that, before I believe that it says what you're saying it says.

Holden: You should read it.

Eliezer: Did the paper say: "And therefore, we should work on reducing on x-risk nowadays as our top philanthropic priority?" Did it actually say that?

Holden: It said, "therefore, maximize expected utility reduces to minimized existential risk."

Eliezer: So that's not necessarily wrong.

Your argument is just that the deworming charity minimizes x-risk. I'm pretty sure that everything you just said reduces to that. I probably should have said that already.

Holden: So certainly not, because again, remember this thing in the whimper, and I guess you're defining whimpers as …

Eliezer: Whimpers are x-risk. They're in the original x-risk paper, right? Maximize probability of okay outcome. From my perspective, the entire thing you just said was an argument for deworming as maxipok.

Luke: I don't see it that way. GiveWell is explicit about the conditions met by the early top recommended charities, and I do wish the GiveWell top charities page said a sentence about how it's not clear to us that this is the maximizing option, but…

Holden: Yeah, that's fair. GiveWell has caveats about we're not doing original research yet and that could be more valuable.

But we do right now, they're all over the place.

Luke: I mean earlier GiveWell.

Holden: Yeah, earlier GiveWell… we really did believe that we and the donors we were serving knew so little, that this was the best thing you could do. Our mental energy was in solving a particular problem and getting people jazzed up about it…

Luke: And part of the decision calculus was: what can we demonstrate knowledge of and get people excited about.

Eliezer: So I thought your entire argument was accepting Maximize Probability of Okay outcome and then making the case for "GiveWell is on the pathway to maximizing…"

Holden: No, no, that was if I accept this ["far future considerations dominate utility" hypothesis].

Eliezer: Yeah, so from my perspective, it's just sort of like I try to have a time-independent utility function and to me, this argument: imagine you were one of these people in the future galaxy, would you like to have never existed? My answer is "no," and then I find that very persuasive, because why should I… things at different times are sort of equally real to my utility function. I don't see why they should be of a different order.

Holden: Let's come back to that. Let me just now state — and you're free to give the opposite criticism of early GiveWell, which I think would be valid — but I just want to say that my experience with the community has been being pointed at this essay, "Astronomical Waste," you read it, you come to your own conclusions. If my picture was some very nuanced picture of existential risk actually includes whimper and this essay actually means that you should reduce the probability of a bad outcome, and blah, blah, blah. It could have been literally true. But this essay read as though t[it] had proven, based on these very speculative calculations that now we should [focus on prevention of direct x-risks such as asteroid impact].

Then I talk to someone like [Person 1], who hadn't read it as "let's make the world better, and figure out what the most important causes are." She had read it as oh, yeah, we need to focus on things that might actually kill everyone that we can think of. That was a pretty bad experience that made me not very, for a long time, not very interested in the community relative to what I could have been otherwise.

Eliezer: Yeah, I'm going to check this. But Nick Bostrom is generally pretty smart and I suspect he might have just have been misread. Like that never happens to me, like that never happens to you?

Holden: No, I know, so take a look and tell me what you think.

Eliezer: (Reading “Astronomical waste.”) “Maximize the probability that colonization will eventually occur…”

Holden: You should read the whole essay.

(Lunch break.)

Eliezer: Okay, I checked the “Astronomical Waste” paper and everything in there seemed correct, but I can see how we would all now wish, in retrospect, that a caveat had been added along the lines of “and in the debate over what to do nowadays, this doesn't mean that explicit x-risk focused charities are the best way to maximize the probability of okay outcome.”

Holden: Right, and in fact, this doesn't tell us very much. This may prove a useful framework, it may prove a useless framework. There’s many things that have been left unanswered, whereas the essay really had a conclusion of: we've narrowed it down from a lot to a little.

Eliezer: I don’t remember that being in that essay. It was just sort of like, this is the criterion by which we should choose between actions, which seems like obviously correct in my own ethical framework.

Holden: I also don’t agree with that, so maybe that's the next topic.

Eliezer: Yeah. Suppose that you accepted Maximized Probability of Okay Outcome, not as a causal model of how the world works, but just as a sort of a determining ethical criterion. Would anything you're doing change?

Holden: I've thought about this, maybe not as hard as I should. I don’t think much would change. I think I would be relatively less interested in direct, short-term suffering stuff. But I'm not sure by a lot. Actually, I think I would be substantially now. I think five years ago, I wouldn't have changed much. I think right now I would be, because I feel like we're becoming better positioned to actually target things, I think I would be a little bit more confident about zeroing in on extreme AI and the far future and all that stuff. And the things that I think matter most to that, but I don't

think it would be a huge change.

Eliezer: Why Extremistan? The entire argument you just gave was precisely why you get to Extremistan eventually without ever passing along Extreme Street.

Holden: I just think there's also a chance that this whole argument is crap and… so there is one guy [at GiveWell] who is definitely representing more the view that we're not going to have any causal impact on all this [far future] stuff and there is suffering going on right now and we should deal with it, and I place some weight on that view. I don't do it the way that you would do it in an expected value framework, where it's like according to this guy, we can save N lives and according to this guy, we could save Q lives and they have very different world models. So therefore, the guy saying N lives wins because N is so much bigger than Q. I don't do the calculation that way. I'm closer to equal weight, right.

Eliezer: Yeah, you're going to have trouble putting that on a firm epistemic foundation but Nick Bostrom has done some work on what he calls parliamentary models of decision-making. I'm not sure Nick Bostrom would endorse their extension to this case, but descriptively, it seems a lot of what we do is sort of like the different things we think might be true get to be voices in our head in proportion to how true they are and then they negotiate with each other. This has the advantage of being robust against Pascal's Mugging-type stuff, which I'd like to once again state for the historical record: I invented that term and not as something that you ought to do! So anyway, it's robust against Pascal's Mugging-type stuff, and it has the disadvantage of plausibly failing the what-if-everyone-did-that test.

Holden: I think the what-if-everyone-did-that test goes really well for it. I think it does better on the "what if everyone did that" test. I think the other way of doing things has much more risks of like [fanaticism]. This test is much more, I don't know, I think it works really well if everyone does it. For example, look at what GiveWell is doing, we're putting … I anticipate putting substantial resources into the highly risky, far future stuff. I also think that like this model of decision-making would be perfectly consistent with doing what you guys are doing, with spending my life on something because I believe in specialization. So I don't really think this creates a problem. I think this is actually much stronger on the what-if-everyone did it, than the more literal expected value, which encourages you to basically take your best,

basically you have the voice that has the biggest number in it and you just follow that one. That has a much easier time justifying stealing money and giving it to charity, for instance.

Eliezer: I'm not sure that speaking… as someone who did try to do principled epistemologies robust to Pascal's Mugging. If you assume that everyone is always mistaken all the time about tail end risks, then obviously an epistemology which ignores all tail end risks all the time is going to do better. The trouble is, what if people aren't always mistaken about tail end risks all the time.

Holden: But if you have different voices in your head and then you negotiate, that doesn't mean you do nothing about tail end risks.

Eliezer: What I'm saying is suppose there were such things as negative lottery tickets? Where even you would believe in the epistemology of them, that it's just like lottery tickets, they just happen every now and then. And everyone pays a dollar for the negative lottery ticket. Well, actually I guess if it literally works like our current lottery tickets, that's fine, because their net payout is less than the dollar everyone gets for it. More along the lines of: there is a lottery ticket which pays you a dollar, and with probability one in a hundred thousand costs someone else a million dollars, and it's well-calibrated, you can see this happening. So in that case, even altruists who obey the "voice in the head gets proportional to probability" principle, will buy all these lottery tickets. So the bad thing is if some low tail end risks, some of the time, are real.

Though that doesn't bother me very much, because I think of AI as a mainline, greater than 50 percent chunk of probability, rather than a tail risk.

Holden: I don't agree with you about the lottery ticket problem. I think partly it has to do with treating robust and non-robust probabilities differently, but I'm not sure that that is the best use of the …

Eliezer: Well, it's kind of unlikely you could blatantly violate the living daylights out of the Neumann-Morgenstern axioms, and not end up with the problem somehow. So in a lot of senses, I see GiveWell as stuck in an unhappy medium between what I think of as normative decision theory and completely naïve charity. The proves too much flaw: if you apply the "attend to probabilities rather than scope" argument, you then sort of withdraw from GiveWell and go back to dumber charity where instead of having

to worry about AMF, you get to see that this person right in front of you is helped.

Holden: Again, it has to do with robust versus non-robust. I think, a lot of my epistemology can just be described as don't do stupid things, whether or not you can formalize them; and don't do things that would lead you to do stupid things.

Eliezer: I explicitly teach this principle as valid, yes.

Holden: Sure. So don't let, when you've got this wild guess speculative stuff, don't let it dominate another worldview that you think is just as plausible, just because the made-up number in it is bigger. But on the other hand, if you're looking at something you actually understand, and your estimates of the scope are more robust, then yeah, let it dominate.

Eliezer: Maybe you already answered this, and my brain just junked the answer, but I think the question of how… so if you're stating that your current strategies wouldn't change very much if you just bought the astronomical stakes, not short-term, but ethical criterion outright, then it does sound like it's okay to interpret the previous part of the transcript as being under maxipok, GiveWell is doing the right thing. Under Maximize Probability of Okay Outcome, GiveWell is doing the right thing.

Holden: No, we're doing something not too far from the right thing. I didn't say it wouldn't change at all.

Eliezer: Right-ish thing?

Holden: Yeah.

Eliezer: Okay, so let's also talk about the N lives thing.

Holden: Okay. So I think I mostly said what I had to say about N lives, and I think it probably might be unsatisfying to you guys. I just think that this argument is too made up and too speculative for me to just buy it and it comes back to that scope thing. I just don't believe it is a good idea to listen to someone say something that's very speculative and has a lot of made up numbers in it, and is about the far future and let that heavily affect your behavior. This is a heuristic.

Eliezer: Can you give me a probability that most of … not a scope-adjusted probability, just a raw probability — that most of the quality-adjusted life-years or utilons, or whatever you want to

call them, ever to be gained, will in fact come from a post-solar future?

Holden: I think there is large probability that I consider that question to be some sort of incoherent or absurd or irrelevant or just ill-formed or unhelpful question. That's where most of my belief comes from. I think to the extent that question makes sense, I think the probability is high, but I also think that question may just be a silly question. This is just the kind of thing, I just don't really believe in grounding much of what I do on abstract philosophy, based on thought experiments.

Eliezer: In my personal, subjective experience, I walk outside at any night and look at up the stars, and I know that the stars aren't little tiny points of light, they're vastly more raw materials than we have access to at the moment. It seems like, I guess, to me it just feels like very straightforward to see how space colonization happens, and sort of silly to think that it's just going to be this one star.

Holden: What about the Fermi Paradox?

Eliezer: I don't know. I don't have any good answers to it, and therefore it doesn't shift my probability estimates much and I just sort of do all my calculations as if the Fermi Paradox wasn't a thing, rather than letting the Fermi Paradox, which I don't understand, be the determining factor in my policy decisions? That sounds an awful lot like things you just said.

Holden: No, I think there's a good outside view reason to believe that the probability of colonizing all those stars is lower than it seems to us.

Eliezer: I don't think that's a good resolution of the Fermi Paradox, although this does get us fairly rapidly into a long conversation. It is very difficult to come up with failure to colonize scenarios that you think are going to apply uniformly across a thousand different intelligent species with a thousand different evolutionary histories and that are also physically plausible. You can't just say nanotech is impossible – it would rule out biology, too.

Holden: What about the amount of energy it takes to find the next one that's colonizable?

Eliezer: If you're taking stars apart, then it's not like you're going around

looking for natural planets.

Holden: What?

Eliezer: Any star is colonizable if what you do to colonize a star is take apart the star.

Holden: Right, so what if it’s not feasible to take apart a star?

Eliezer: Then we must have learned some amazing new fact about physics. Why can’t you build a Dyson Sphere? It’s the wrong kind of amazing new fact about physics, it’s not like in the 12th decimal place, it’s not a new fundamental force, it’s: you are bizarrely and magically prohibited from using the physics we know to construct these sort of very straightforward …

Holden: No, not bizarrely and magically, it just turns out not to be doable.

Eliezer: Okay, so I guess I put a lot of faith and credit in the arguments I've seen that if the higher level than quantum laws of physics are what we think they are, we can just do this stuff.

Holden: Let me step back a second. I hear your claim that I should assign a very high probability that we can — if we survive — colonize the stars. I believe this to be something that smart technical people would not agree with. I've outlined why I think they wouldn't agree with it, but not done a great job with it and that's something that I'd be happy to think more about and talk more about.

Eliezer: Are there reasons apart from the Fermi Paradox?

Holden: I don't know what all the reasons are. I've given my loose impression and it’s not something that I've looked into much, because I didn't really think there was anyone on the other side.

Eliezer: Yeah, I think that we have sort of different commonsense notions here. From my perspective, the notion that we're going to get most of your QALYs from the future is around as commonsensical as the notion that most of your food would not come from any one restaurant you picked, because there’s a lot of restaurants, there’s a lot of stars.

Holden: Well, one of them has a bunch of actual facts in support of it, and the other is kind of this logical argument that may have a lot of baked in assumptions that are wrong. I actually don’t get most of

my food from one restaurant. That's a huge difference between the arguments.

Eliezer: I mean, argument by analogy is not going to settle this, but I just wanted to state that I think that down to earthness is this major epistemic virtue. It seems very down to earth that most of your utilitons don’t come from down on earth.

Holden: I think this is actually … this is not something that I think of as a matter of common sense, this is something that I would easily change my mind on if I thought that the people who knew about it were in a certain camp and my impression is that they're actually more in my camp, that I'm describing right now.

Eliezer: If it comes to that, I can get [these people] to endorse “most of your utilons come from other stars.”

Holden: Okay, that's cool.

Eliezer: We can try that.

Holden: That's interesting. I think that would be a good … I mean, that is probably a more productive path to getting me to endorse it.

Eliezer: Some amount of uncertainty in that estimate because you've obviously spent more time around these people than I have.

Holden: Yeah, but not a ton. This isn't something that I put a ton of thought into.

Eliezer: But it feels to me both physical common sense and not very far from the zeitgeist I expect they grew up with. So we can test this.

Holden: Sure. Sure.

Luke: I wonder if anyone has ever done a survey.

Holden: Yeah.

Eliezer: Of that exact question?

Holden: Yeah.

Luke: Something pretty close to that.

Eliezer: I would probably want to condition the question on: no materialized existential risk, because if existential risk materializes, then you get most of your utilons from present day.

Holden: No, we'd have to make the question carefully written, obviously.

Eliezer: But just sort of like the common sense version, if all goes well, not super well, just moderately well or something like that.

Holden: Yeah.

Luke: I guess the people I think of immediately are people at NASA and Stephen Hawking-type who say…

Eliezer: That's kind of selected.

Luke: Right, and the problem is that a very decent explanation of why they're saying we need to hurry up and colonize the stars is because that's their funding.

Holden: Well, let me make another comment on N lives, and I don't know how you're going to respond to this, and I haven't even discussed this with a lot of people, but this is on the other side of N lives. This is on the: does creating a life count the same as saving a life. So I'm not sure, again, if there's a multiplier, I think the multiplier is high enough to not wipe out the big number. But I'm not sure the multiplier framework works. So one crazy analogy to how my morality might turn out to work, and the big point here is I don't know how my morality works, is we have a painting and the painting is very beautiful. There is some crap on the painting. Would I like the crap cleaned up? Yes, very much. That's like the suffering that's in the world today. Then there is making more of the painting, that's just a strange function. My utility with the size of the painting, it's just like a strange and complicated function. It may go up in any kind of reasonable term that I can actually foresee, but flatten out, at some point. So to see the world as like a painting and my utility of it is that, I think that is somewhat of an analogy to how my morality may work, that it's not like there is this linear multiplier and the multiplier is one thing or another thing. It's: starting to talk about billions of future generations is just like going so far outside of where my morality has ever been stress-tested. I don't how it would respond. I actually suspect that it would flatten out the same way as with the painting.

Eliezer: I often suspect that I may be an average utilitarian because the numbers involved are so large that I can't aggregate them properly and have to start thinking of them in terms of fractions. But then I'm like an average utilitarian over the multiverse, not just my section of it. So if I think I'm above average for the

multiverse, I want more people to exist here. I suspect I may end up, either neutral or favoring use of resources for existing people, rather than creating new people, but that's only true because I think you can get at least as much utility out of using more resources for existing people, rather than creating new people.

Whereas those new people that have lives that seem like very worth celebrating, and so if you can do better by using more resources for existing people, that must be even more worth celebrating. I can imagine a future in which there is seven billion people who are alive at the intelligence of explosion, and are using all the galaxies, or maybe like one galaxy per person, so just like a hundred billion people. But if so, that's because that alternative will actually be better than having a quadrillion people, or a septillion people or something like that. Because those very large entities were able to contain more happiness than if we took the same amount of computing power and distributed it over lots of other entities.

So to me, it feels like visualizing a happy intergalactic civilization and celebrating all of its good deeds and happiness, is a lower bound on what it’s worth.

Holden: I agree with that. But I just think that it’s very possible to me that as we add lives right now, to a happy world, and they’re happy lives, and valuing each one at I don't know, saving a life times point two or something. And simultaneously, my morality is approximated by utilitarianism in most of the cases that I confront. Then simultaneously, a happy civilization of N people is three times as valuable as a happy civilization of seven billion people [even when N is much more than 3x 7 billion].

Eliezer: Suppose that there was some sort of relatively painless ailment that killed people like heart attacks. We can trace the heart attacks back to a certain chemical that's now very common in the environment. Suppose that somebody in the past, suppose they had known this in the past and someone was, “Well, sure, by cleaning up this chemical, we can give people in the future longer lives.” But increasing lifespan like that seems to me making a larger painting.

Holden: Yeah, I'm not sure what to say to this. A, we haven't increased lifespans that much, we've doubled them. B, I would probably disagree with that person, but I don't know if I would disagree with them if we multiplied it by a hundred. I'm also not totally

sure I would disagree with that person. C, in general, you're just not going to get me to change of mind with thought experiments, very often. I'm not saying it will never happen, but especially thought experiments that really hinge on things that are so exotic, I just don't really …

Eliezer: Interesting. The point I was aiming toward there was to set up: if you value aggregated lives, then the result of extending lifespan must be at least that good, and then prove to you that you cared about extending lifespan, or something along those lines. If you don't care about extending lifespans, the endorsement of deworming is really hard to explain.

Holden: No, deworming improves the quality of life.

Eliezer: Okay, interesting.

Holden: I also greatly value the possibility of a future happy civilization. I don't need a linear valuing of each extension of year the same, in order to believe that something roughly a thousand times the population we have and going from wherever we are to ten out of ten on the goodness scale would be really awesome and would be awesomer than saving all the lives that are there today. I can accept all that, but not believe that it scales all the way to where these N sets are taken literally.

Eliezer: What if our civilization was one tenth the size, do you feel like we'd lose much less than 90 percent of the utility?

Holden: I'm just not sure.

Luke: So I'll say that I also find myself to have significant normative uncertainty about what my values are, and the way I think about it is that I have a value system at age ten or whatever that is produced with almost no deliberation. It's just produced by my family and evolution and stuff like that. So the reason I value thought experiments and learning more about the world, like block universe or whatever, is that it doesn't resolve the issue, but it makes my values more a product of trying to think through them and test against that thought experiment, that thought experiment …

Holden: Thought experiments make me think, but my answer is often "A and not-A and I reject A contradicts not-A." That's often my conclusion and that's why thought experiments don't work all that well on me. It's not always my conclusion.

Eliezer: Does that mean that if I convince you of the principle A is A, I can get you to support x-risk?

(Laughter.)

Holden: (Laughing) Where A and A are things that sound the same, but then I'm not sure are actually the same.

Luke: Right, sure.

So what's a desirable process for figuring out what you should value?

Holden: Thought experiments are [helpful]. Learning more about the world and talking to people is [helpful]. These things sometimes change values. They don't change values according to a predictable algorithm and I largely feel I expect to remain uncertain.

Luke: One way I interpret our difference here might be, and you might feel this is uncharitable, so feel free to correct me, but… I was having a similar conversation with someone else, who didn't seem very interested in changing their values as a result of new information or new thought experiments, and so on. It was sort of like I had just a weaker prior on the values I happened to have at age ten. This other person had almost like a stronger prior, and I actually convinced them that they should have a weak prior on epistemic grounds, but then they just said, okay, it's not that I'm confident about my values, it's that I'm gung-ho about them, irregardless of my epistemic confidence…

Holden: That's not really where I am. I'm not confident about my values at all. It just doesn't …

Luke: You just don't think you can …

Eliezer: It seems to be there is a bit of analogy here. The notion of GiveWell as being halfway between maxipok and normal charity. So at the EA Summit, you said something in response to an x-risk question, you said something along the lines of, "Well, it feels to me like these astronomical stakes can be an excuse not to evaluate things like AMF."

Holden: Right.

Eliezer: I now have a better idea of why you would say something like that, although at the time, I was rather shocked. But it seems to

me that the sort of corresponding argument, someone is “well, why give to deworming instead of the local classical symphony?” You sort of try to convince them to be a bit more of an aggregative utilitarian, and then they may reject thought experiments. It seems to me that, an exactly analogous argument for we're not going from …

Holden: I can completely see that, yeah. No, I totally see it. To me, there is a threshold of argument strength that changes my mind and then there is another kind of argument that doesn't change my mind. I've landed where I've landed and the fact that there is a certain formality with which I can’t prove it, doesn't really change my mind about where I am. I have become convinced that the symphony thing is not right. I have not become convinced that the N lives thing is right. That's where I stand and [while I'd ideally be able to formalize it better] I guess I don’t really see why that's a problem.

Eliezer: Okay.

Holden: Basically, I understand that my thought process, if you fed different parameters [and experiences] into it, could lead to someone validating the symphony and that doesn't bother me very much. My process with the parameters that were fed in, generated what it generated, I'm following it where it goes and finding people who resonate with me enough to help push us forward. And that's that. I'm not totally sure those classical symphony people are wrong, and I'm not totally sure you guys are wrong, but I'm where it seems right.

Eliezer: Not sure that thought process is going to be very accessible to third parties who are trying to make similar decisions themselves.

Holden: Well, some of it isn't, and that's just the nature of human communication, is that there is always a bunch of stuff below the surface that is kind of assumed, and I try to make it explicit, and I try to explain why, and if I were to say what's the difference between classical symphony to deworming and deworming to existential risk, I would say that there is more stuff, there is more data points. It’s not just thought experiments. I've been to Africa, I've looked around, I would encourage people to do that.

Eliezer: If I could just take you to the future, this would be so much simpler!

Holden: I get it, I get it. But there’s more stuff there. There’s more data on Africa and that matters. It’s not just having been to Africa, it’s a whole bunch of stuff, and some of it is expert opinion, too, that influences me on these questions I'm highly uncertain about.

Eliezer: Would you endorse the statement that if you can predict how your opinion will change, you ought to predict, you ought to change in that direction?

Holden: Yeah.

Eliezer: So if you could visit the future …

Holden: Well, actually … (Laughs) Well, I've certainly imagined visiting the future. I don’t actually fully endorse that statement, because I also think that there is different cognitive processes running in my head and negotiating with each other, like we've described. Sometimes I see value in separating Holden, who tries to be mostly normal and meet normal standards, and Holden, who is trying something crazy that just might work. Those two often believe different things and respond to the same thing differently.

Eliezer: I think the best common sense version of what I'm doing is something along the lines of: it’s not even based on any deep moral principles, as the sense that if I could actually see and comprehend the intergalactic future, I'd feel like, “of bloody course that was a very large number of utilons!” It’s the predictable update if I can see it, that's driving me here.

Holden: I don’t use that process as much as you.

Eliezer: Ah. I think this is a pretty good place to stop … do you think so?

Holden: Let me say one other thing. With a lot of this stuff I feel like… I hear an argument and I think, "That's probably wrong for reasons I just can't think of right now. And that's where I am with a lot of this stuff. And that's a function of who is saying it, how credible I think they are, what their track record is, what the intellectual methodology is, how good its track record is, and a bunch of outside views and heuristics that make me say "Ten years from now I probably will see the problem with this."

Eliezer: So my argument for "If you could visit the future you'd think of course it's more valuable" sounds reasonable, but you think that in the future you'll know how to refute it.

Holden: Yeah, exactly.

(Laughter.)