EA Global Bay Area: 2024 | Cause Prioritization & Global Catastrophic Risk | Hayley Clatterbuck

Talk available at https://youtu.be/AkGHvUDW8q4

Model available at https://ccm.rethinkpriorities.org
(browser-based, so you can use it on your phone)

Glossary (in rough order of appearance):

- Cause Prioritization
    - the process of researching which charitable causes offer the most benefit given marginal investment

  Global Catastrophic Risk (GCR)
    - refers to a risk that could inflict "serious damage to human wellbeing on a global scale" without necessarily causing extinction
- Existential Risk (X-Risk)
    - The risk of curtailment of the potential moral value in the future (e.g. many many future humans), either by extinction or unrecoverable civilizational collapse.
- Effective Altruism (EA)
    - The practice of trying to figure out how to do the most good with the resources at your disposal and then carrying out the answer you arrived at.
- Cause Areas
    - E.g. global poverty, global health, animal welfare, catastrophic risks - problems to be solved or populations to be helped. This talk references the cause areas most popular within the effective altruism community (global health, animal welfare, existential risk)
- Expected Value (EV)
    - The average payoff of an action, calculated by multiplying the probability of each possible outcome by the payoff in that outcome. E.g. buying lottery tickets has negative expected value
- (Moral) Value
    - When we say "the value of the future" we mean the moral weight of, or the moral consideration owed to, beings that may exist in the future.
- Longtermism
    - The idea that most moral value may exist in the future because there could be many many generations of humans and/or morally relevant beings in the future (e.g. trillions of future people).
- AMF (Against Malaria Foundation)
    - A charity that distributes mosquito-killing bednets in regions impacted by malaria.
- Welfare ranges
    - Estimates of the differences in the possible intensities of animals' pleasures and pains relative to humans' pleasures and pains. An input in to the moral worth of a being, e.g. if a shrimp has a limited capacity to have positive or negative experiences it has a correspondingly small welfare range and proportional moral consideration (depending on philosophical credences)
- DALY (Disability-Adjusted Life Year)
    - One DALY represents the loss of the equivalent of one year of full health. DALYs for a disease or health condition are the sum of the years of life lost due to premature mortality and the years lived with a disability (valued at some lower fraction based on the disability) due to a condition.
    - In the EA community we may say something like "the cost-effectiveness of this intervention is 1 DALY for $1000", by which we mean 1 DALY is averted or the equivalent of saving someone's life for one year.

YouTube Transcript (lightly cleaned up + slides)

Thank you for coming out on this rainy morning. [banter]

Anyway, it's great to be here. So as Derek mentioned, in the fall of last year,
our team at Rethink Priorities, the Worldview Investigations Team undertook a sequence of research that we called the CURVE [Causes and Uncertainty: Rethinking Value in Expectation] Sequence.

And the sort of stated goal of this project was to take various tools of evaluating the cost-effectiveness of certain actions across different cause areas.

And so we were comparing things like existential risk mitigation efforts against animal welfare efforts against global health.

So what I wanna do in this talk is first introduce you to some of the work that we undertook there, and then secondly, to extract some of the lessons that we learned from making, sort of, cross-cause comparisons to see if we get any insights about how to do cause prioritization within the GCR space itself.

So in this project, we took on what we take to be two influential ideas in EA.
We think that these ideas are both commonly held and have been really influential in moving a lot of money in the last few years. So we wanted to give them a deeper look.

So these two ideas are that

1) expected value maximization supports prioritizing existential risk mitigation work

over all else - so that when we do the cost-effective estimates this way, we find that existential risk mitigation is robustly favored over global health and animal welfare causes. In this talk I'll largely be talking about x-risk [existential risk] because that's the way we phrased it in CURVE and occasionally, sort of, expanding to a broader GCR framework.

2) The second main platform of our research sequence was questioning whether expected value maximization is how we ought to be making decisions in the cause prioritization space.

So on this first component, we showed that the expected value of x-risk mitigation efforts significantly depends on a range of assumptions that are really difficult to assess.

So depending on how you come down on these assumptions, sort of what you set as, sort of, your parameter values for certain key variables, you can get x-risk mitigation efforts to be either orders of magnitude more cost-effective than other kinds of cause areas to about equal or even less cost-effective.

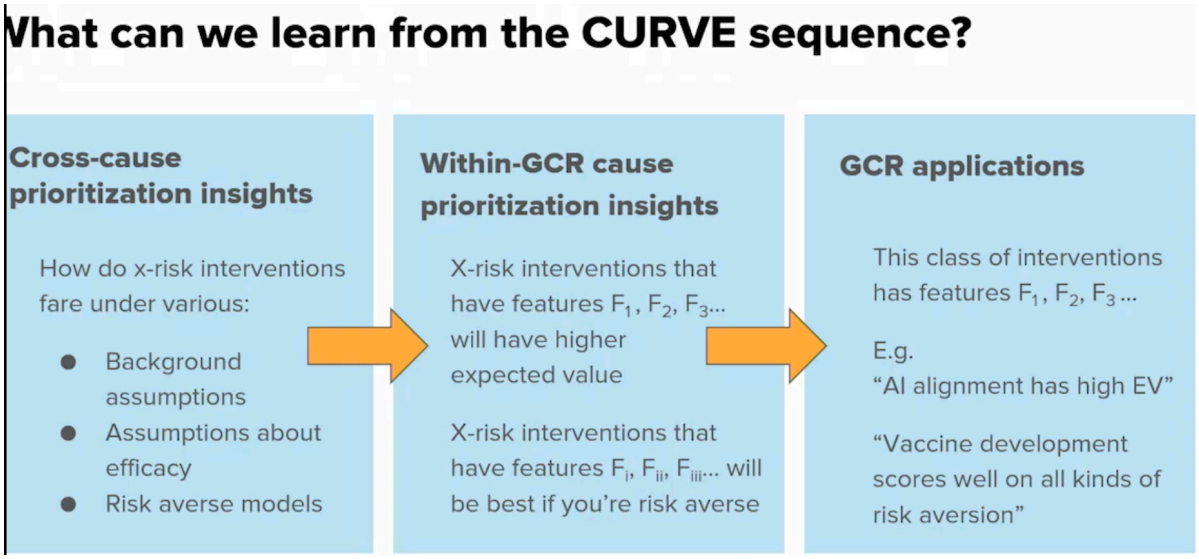
So part of what I'll do today is run through what some of these key variables are and what they can tell us about when x-risk mitigation efforts are better and worse by the lights of EVM [expected value maximization].

Secondly, we looked at some alternatives to maximizing expected value.

So we take it that there are several different kinds of potentially rational risk aversion. And then if you have risk-averse attitudes, you'll differ in significant ways from if you were an expected value maximizer.

Philosophers and economists have developed formal decision theories incorporating some amount of risk aversion, and in some of our reports, we applied these different kinds of models to cost-effectiveness estimates to see what difference they made.

We found that risk-averse models can significantly make significantly different recommendations from EVM, sometimes more robustly in favor of existential risk mitigation and sometimes favoring global health and other kinds of causes.



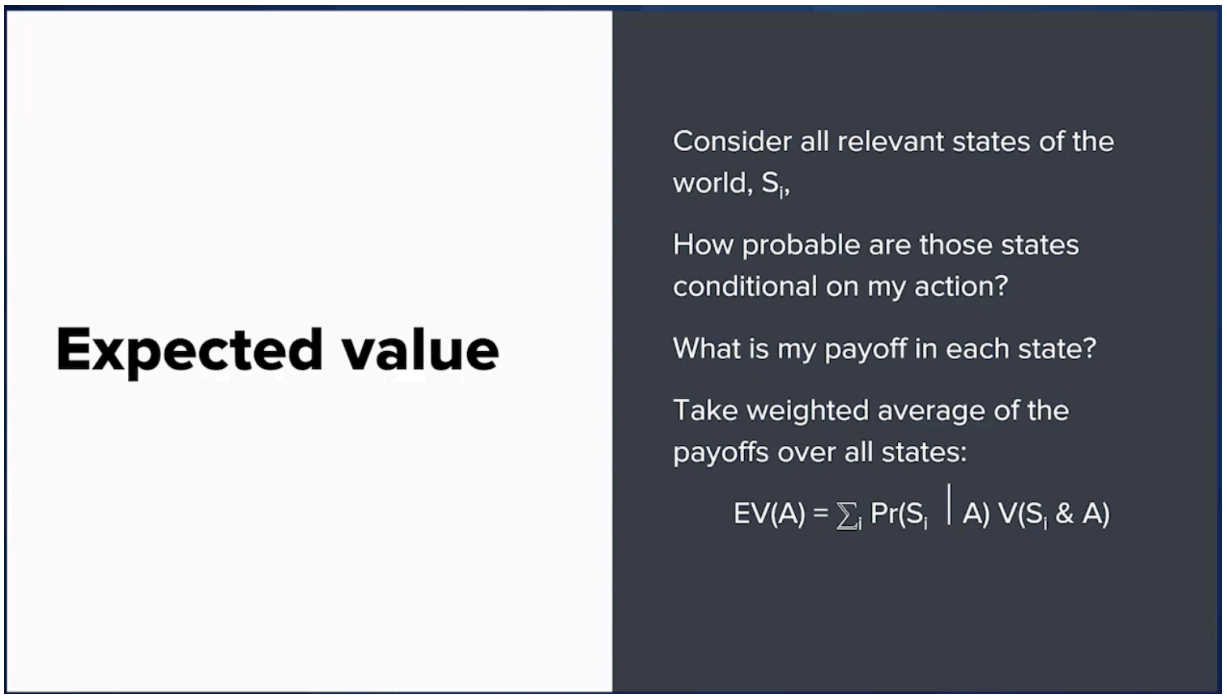
So during this process, we sort of uncovered a lot of information about when existential risk mitigation efforts are better and worse by the lights of various decision theories. So we looked at, how do x-risk interventions vary under various assumptions about the background state of the world, what the world is like, about the efficacy of x-risk mitigation efforts, and about how risk aversion might affect your decision-making?

In this talk, I wanna extract some of the insights that we found there to find a series of features that x-risk mitigation efforts can have that make them better or worse.

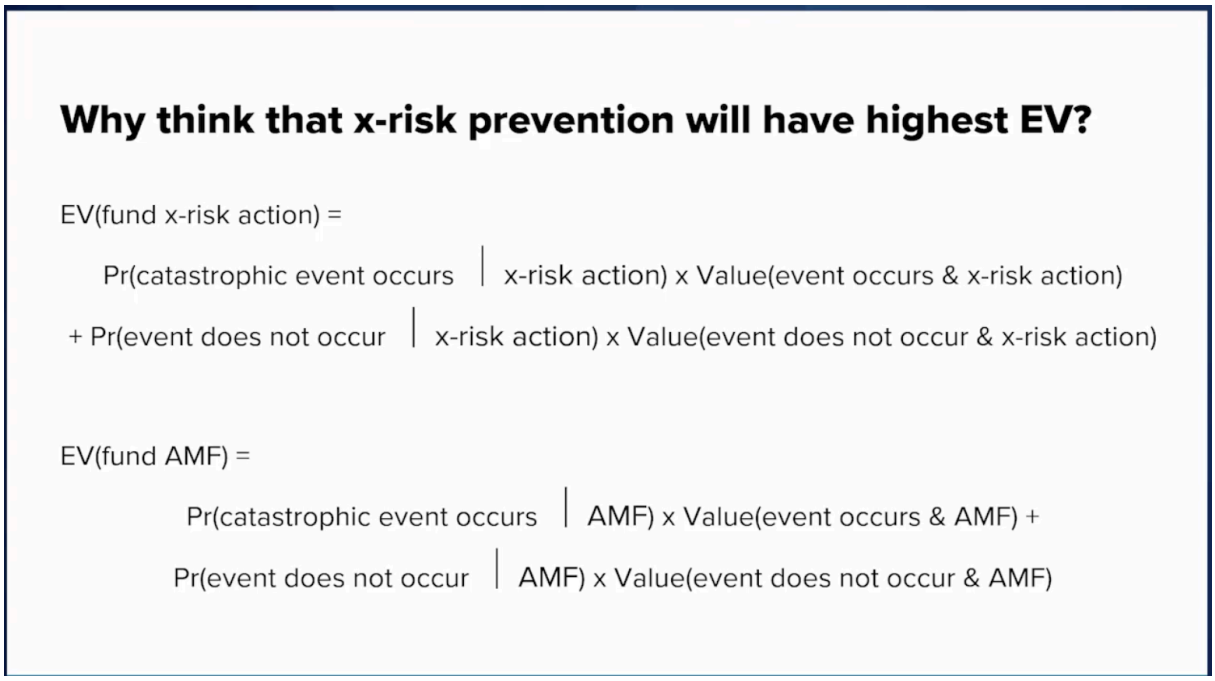
So to find insights like x-risk interventions that have features one, two, and three will tend to have high EV or tend to do well under various kinds of risk aversion. Lastly, you could use those insights to make some decisions within the GCR cause prioritization space itself. So you might identify particular GCR actions that have those admirable features that we uncovered. I will have a lot less to say about this third, sort of, component here and in part because that just requires a lot of very specific information and expertise about what GCR actions are available to us and the features that they have. As a philosopher, I'm gonna speak to the number two in a little bit more depth and sort of rely on the expertise in the room at this conference in general for filling out the third thing.

At some points, I will surmise, but I will flag when those are sort of mirror, you know, suggestions that you should take with a grain of salt rather than something that our team has actually investigated.

So let's start with this first aspect of the project, evaluating when do x-risk interventions have high expected value, when do they have low ones? So I'm gonna do the briefest, sort of, introduction to what expected value maximization is for those of us who don't work with it frequently.



So if you wanna calculate the expected value of some action, you consider all of the states of the world that are relevant to making that decision, the things that might matter for the purposes of your action. And then you ask, how probable is it that those states will obtain if I do my action? Then you ask, in each of those states of the world that might obtain, what is my payoff? How good or bad is it for that state of the world to come about? And then to calculate the expected value, you take the weighted average of those payoffs over the states, and that's weighted by their probability of obtaining, given that you act. And a sort of long-standing decision theory that's really well defended says that you should perform an action only if it has at least as high of an expected utility as any other action.



So sort of prima facie, when you sort of first look at it, there are strong intuitive reasons to think that x-risk mitigation efforts are going to have a higher expected value than any other kind of thing that we might pursue. These intuitions sort of underlie long-termist, sort of, positions.

So why might you think this? So here's one way to think about the expected value of an x-risk action versus an AMF action.

And I thought that things would be closer to my arms when I wrote this, so I'll just gesture in the general direction [pointing at the slides].

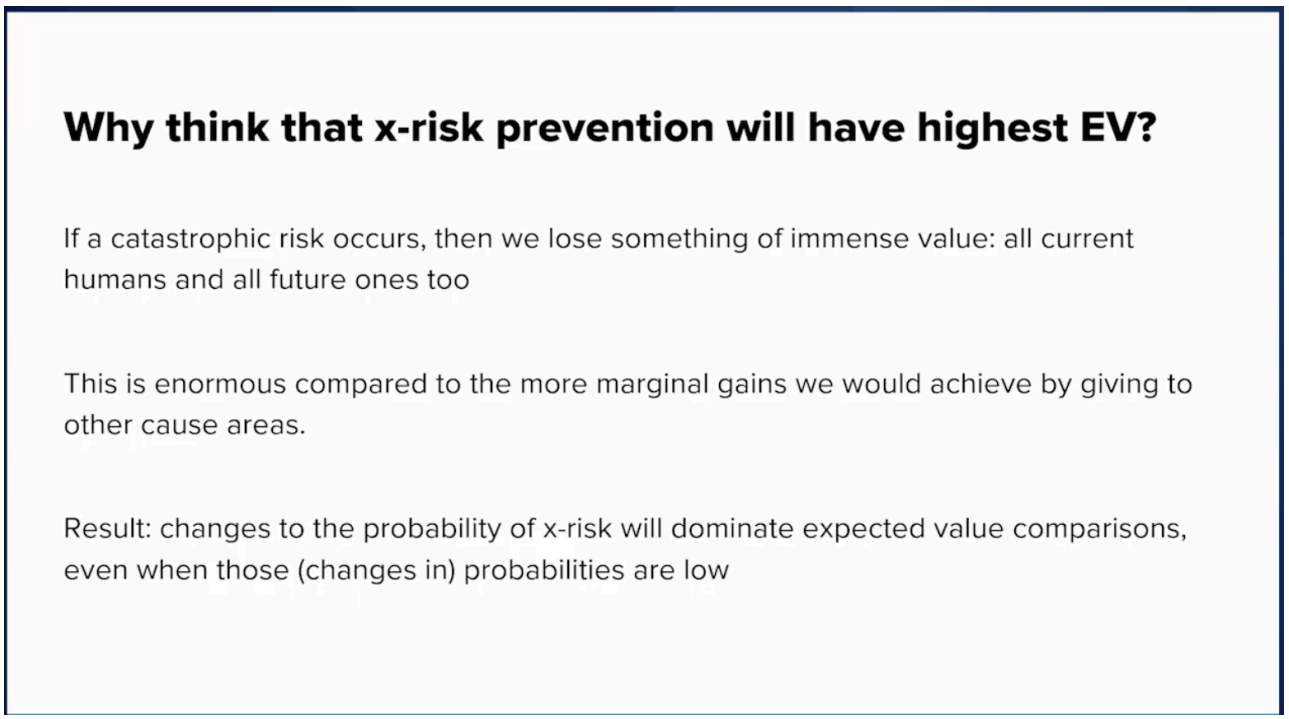
So to calculate the EV of an x-risk action, you say there's some catastrophic risk that we're trying to avoid. So we calculate the probability that that catastrophic event will occur, given that we did our x-risk action, times the value that would obtain if that catastrophic event occurs and we undertook our x-risk action, and then we add that to the probability that the catastrophic event doesn't occur, given that we took our action, and the value that we get if we avoid that catastrophic event and invested in that action. Compare that to the expected value of funding something like giving a considerable, or a comparable amount of money to the Against Malaria Foundation. There we have the probability that a catastrophic event occurs, given that we gave to AMF, times the value that would obtain in that event, and so on.

So notice that the value that we get if we suffer a catastrophic event is so, so, so, so, so low. That's a huge loss. And the value that we stand to gain by avoiding a catastrophic event is extremely large. Given that and given that those are so much larger amounts of value than the marginal gains you could get by giving to AMF, it seems like your expected value calculation is going to be heavily influenced by that astronomical amount of value that depends on whether or not we suffer a catastrophic event.

Furthermore, given that, you might think that the probability that our action, the change in the probability that our action yields for that amount of astronomical value is going to be of the utmost importance.

Probably giving to AMF isn't gonna change that probability of a catastrophic event at all,
but we would hope that our x-risk mitigation action would.

So here's the, sort of, intuitive case, why I think a lot of people have found it intuitive that x-risk is gonna come out ahead.

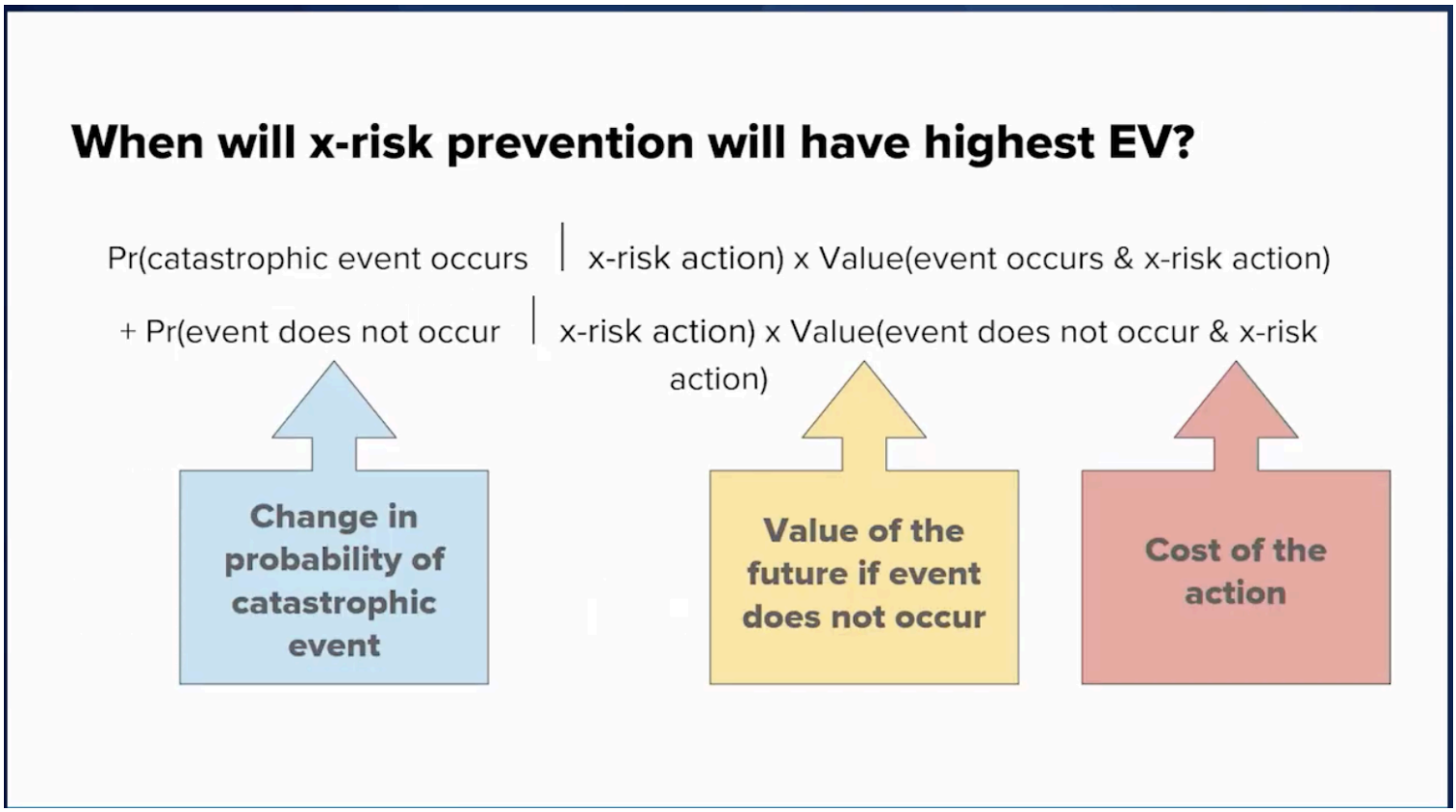


If a catastrophic risk occurs, we lose something of immense value. All current humans and all future ones too. That's enormous compared to the more marginal gains we'd achieve by giving to other cause areas.

So the result is that the changes to the probability of x-risk will dominate expected value comparisons, even when the changes in probabilities they yield are low.

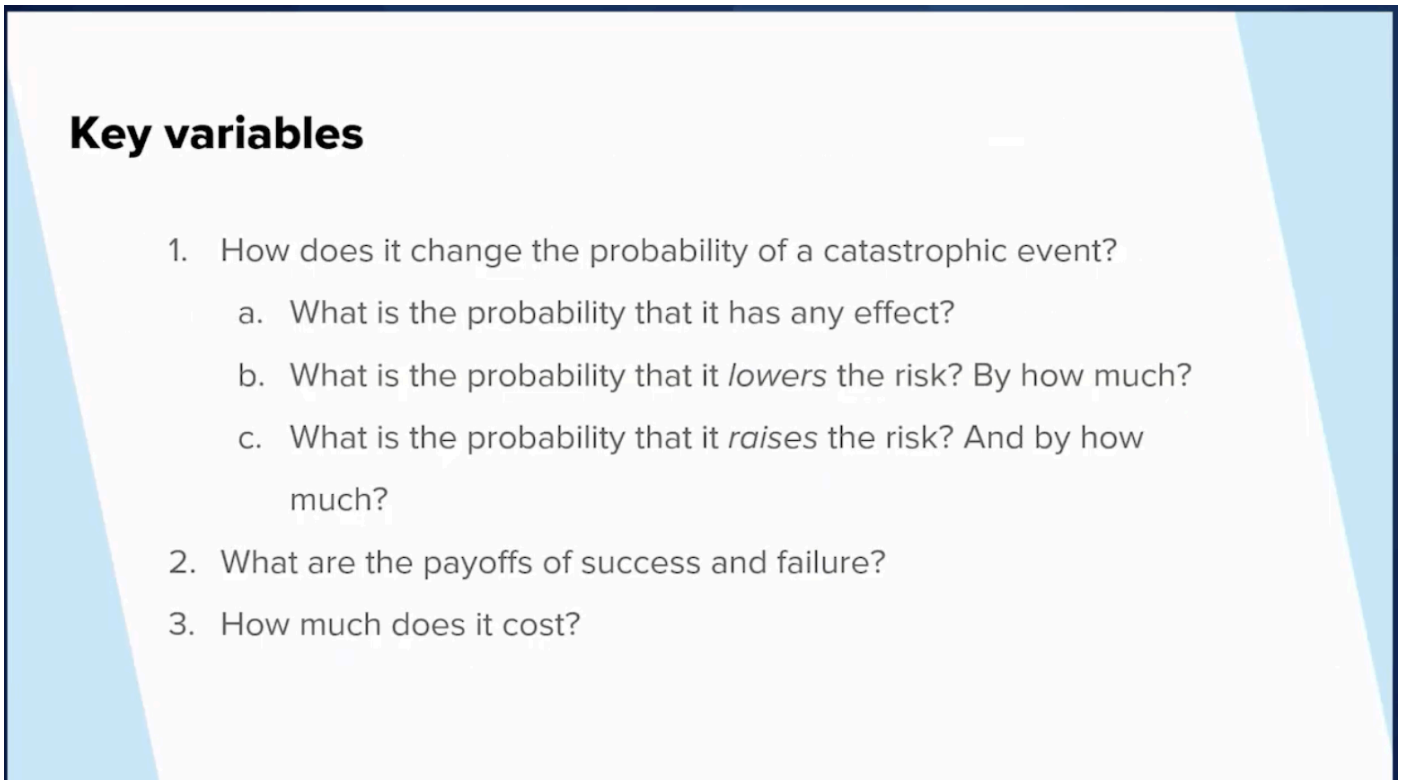
So that's the intuitive case, and I'm gonna give you some information about some projects that we're looking further into this, sort of interrogating this in a little bit more detail.

So if we look again at this expected value calculation for an x-risk event, sort of, there are three components that are really important here.



- One is what's the change in the probability of a catastrophic event, conditional on us doing this action?
- The second is, what is the value of the future if the event does not occur? What's the payoff if we manage to avoid a particular catastrophic event?
- And lastly, you wanna consider the cost of doing the action. That's gonna factor into this cost-effectiveness estimate as well.

So here are the key variables.



First, we can decompose the change in the probability of a catastrophic event into these following sub-components.

- First, what is the probability that our action has any effect on global catastrophic risk. What's the probability that we make any change at all to the amount of risk that exists in the world?
- Secondly, given that we have some effect, what's the probability that we've lowered the risk,

and by how much?

- And thirdly, what's the probability that if we had an effect, it actually raises the risk? What's the probability that our action backfired and made a catastrophic event more likely?

Secondly, what are the payoffs of success and failure of avoiding a catastrophic risk?
And lastly, how much do these actions cost?

So the first project that I'm going to present to you focuses on this first bit, how much changing the probability of an existential risk matters.

And we're gonna do that by making some simplifying assumptions that'll allow us to, sort of, more effectively hone in on one, and then later, I will talk about a project that takes on two.

So the first project I'm gonna talk about is by Laura Duffy, the common sense case for x-risk work.

So in order to zoom in on the probabilities at play, Laura made some simplifying assumptions.
So it's often assumed that even if we don't look a million years into the future, even if we focus on the next two to three generations, we find that x-risk work is more cost-effective than other kinds of projects. So this is what she asked.

Common-Sense Case for X-Risk Work

"Is x-risk the most valuable cause if we just look at the next few generations?"

Premises:

1. Value of x-risk mitigation limited to humans
2. Average annual value of civilization roughly the same as now
3. Only considering next 120 to 180 years

Is this the most valuable cause if we only look at the next three generations, considering only the next 120 to 180 years and assuming that humanity exists in roughly the same numbers with roughly the same amount of value per individual as now? What can we say then?

Existential Risk Reduction Scenarios

	P(Effect)	P(Good | Effect)	Rel. Magnitude of Bad Effect	BP Risk Reduced per $1B | Good Impact	E(BP Reduced/$1B)
Scenario 1	0.2	0.55	0.1 to 1.0	0.5 to 5.0	0.2
Scenario 2	0.2	0.55	0.1 to 1	1.0 to 10.0	0.3
Scenario 3	0.9	0.55	0.1 to 0.6	0.5 to 5.0	0.8
Scenario 4	0.9	0.90	0.1 to 0.6	0.5 to 5.0	1.6
Scenario 5	0.9	0.55	0.1 to 0.6	1.0 to 10.0	1.5
Scenario 6	0.9	0.90	0.1 to 0.6	1.0 to 10.0	3.2
Scenario 7	0.9	0.90	0.1 to 0.6	5.0 to 15.0	7.2
Scenario 8	0.9	0.55	0.1 to 0.6	5.0 to 15.0	3.5
Scenario 9	0.2	0.55	0.1 to 1.0	50.0 to 150.0	6.9
Scenario 10	0.9	0.90	0.1 to 0.6	50.0 to 150.0	72
Scenario 11	0.9	0.90	0.1 to 0.6	500.0 to 1000.0	567
Scenario 12	0.2	0.55	0.1 to 1.0	500.0 to 1000.0	55

Surveys of x-risk researchers typically say an effective project mitigates 0.1 to 10 basis points per billion

14

So this is a big chart, but I'll walk you through what some of it means.

She tried out different, sort of, parameter settings or different, sort of, ways of making assumptions about those key variables that I laid out a few slides ago.

So a green coloring means that she's made an assumption that's favorable to x-risk work, and a yellow one means it's one that's less favorable.

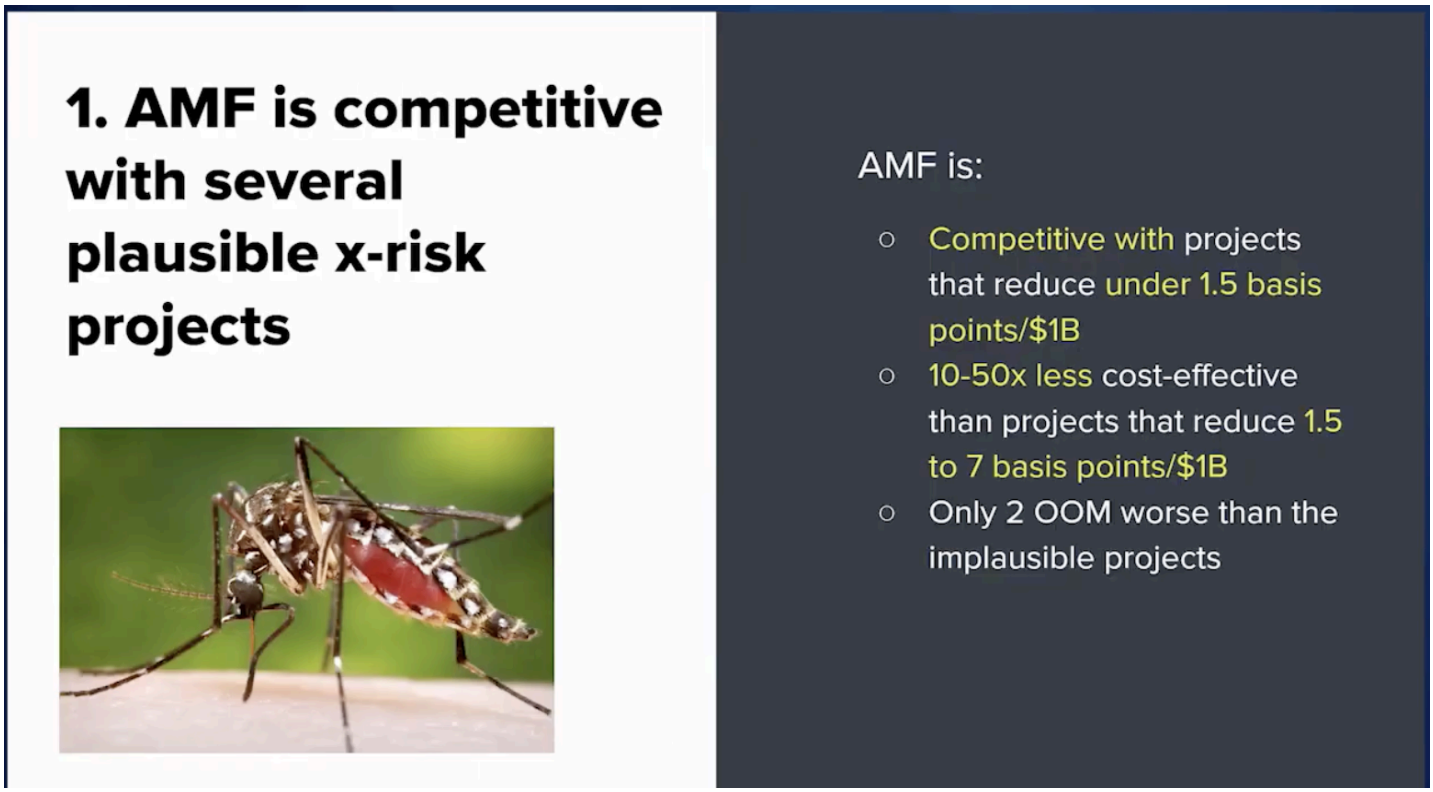
So in the first scenario, for example, you assume that the probability or action has any effect is 20%,
the probability that has a good effect, given that it has an effect, is 55%.

Then you can ask, what's the relative magnitude of a bad effect versus a good one? So she assumes here that if it has a bad effect, it'll be smaller than the potential good effect that the action would have.

And then these last two columns are extremely important as well. So it's the estimate of how many basis points of risk were reduced per billion dollars, given that we assume our action had a good impact.

As Laura points out, surveys of x-risk researchers typically say that an effective project mitigates between 0.1 to 10 basis points [one basis point is 0.01%] per billion dollars. And then she uses that finally to calculate the overall expected basis points reduced per billion, taking into account the other variables
that she's considered. So she modeled how changes to various assumptions here, less or more favorable to x-risk, less or more plausible, make a difference to overall cost-effectiveness estimates.

Here's what she found



AMF is competitive with several of those plausible x-risk projects if we make less audacious assumptions. So for projects that reduce under 1.5 basis points per billion dollars, AMF is roughly comparable to x-risk work.

But for actions that reduce much higher numbers of basis points per billion dollars, those are 10 to 50 times more cost-effective than AMF, even if we just look over the next 180 years.

But even if we make the most favorable assumptions to x-risk work, making some very implausible assumptions in the process, those only turn out to be two orders of magnitude more cost-effective than AMF.

So the intuitive idea that, like, x-risk work is gonna have astronomically more value than AMF just doesn't look to be the case, and they look on roughly a par, given some plausible assumptions we might make.

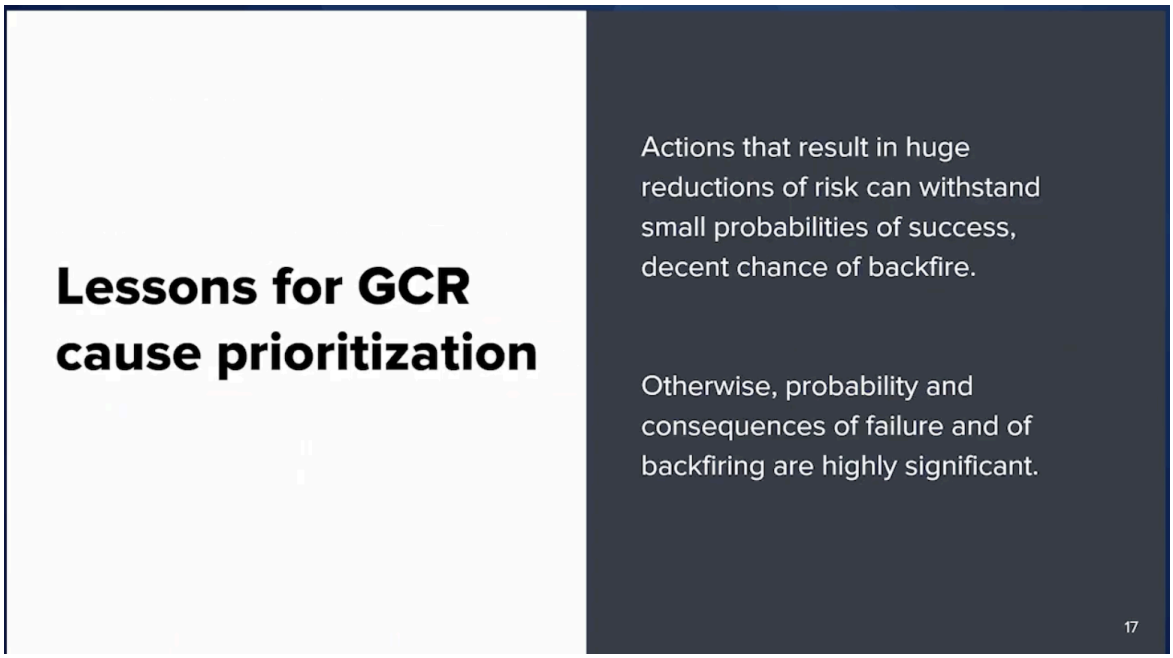
How about animals?

If you use Rethink Priorities' welfare ranges, we get that cage-free campaign to help chickens are at least as good as x-risk in all plausible scenarios.

You need to make extremely implausible assumptions about the value of x-risk work to get them to be more valuable than chickens. And the reason is the number of chickens you can affect with an action is so large that it starts to approximate the amount of astronomical value you might expect from x-risk work.

You might not buy Rethink Priorities' welfare ranges, you might think that they're too high, fair enough.
But even if you assume lower welfare ranges, you can see that only a couple of scenarios are much better than cage-free campaigns.

So what are the lessons that we should learn from this work here?



One is that one of the key variables that seem to be driving a lot of these comparisons is that basis-points-per-billion-dollars variable.

So actions that are expected to have huge reductions of risk, I mean, huge is relative here because a basis point is a percent of a percent.

But things that have large reductions to the probability of a cataclysmic event can withstand small probabilities of success and a decent chance of backfire.

So once you're above a certain threshold, we found that seemed to be quite important.

Otherwise, for more realistic estimates of how many basis points per billion dollars you can avoid, the probability and consequences of failure and of backfiring are extremely important.

That's gonna be one of the themes of this talk is that when you're thinking about an x-risk intervention, you can't just focus on the potential upsides. You need to also focus on the probability of backfiring, of actually making it more probable that a catastrophic event will occur. Backfire becomes extremely important when we're talking about lower amounts of basis points per billion.

So I'm sure a lot of you are thinking you've really hamstrung the case for x-risk work by only focusing on the next three generations. After all, my intuitions were being guided by the potential astronomical value of the future.

So the second report that I'm going to tell you about focuses more squarely on that aspect of it.

How valuable should you expect the long-run future to be? So our colleague, Arvo Muñoz Morán, attacked this question.

So here's a question.

If we prevent an existential event, sorry, an extinction-level event, how much value do we get?

This is gonna depend on two things.
One is how many years long is the future expected to be?
Secondly, how much value per year is there going to be in that future?

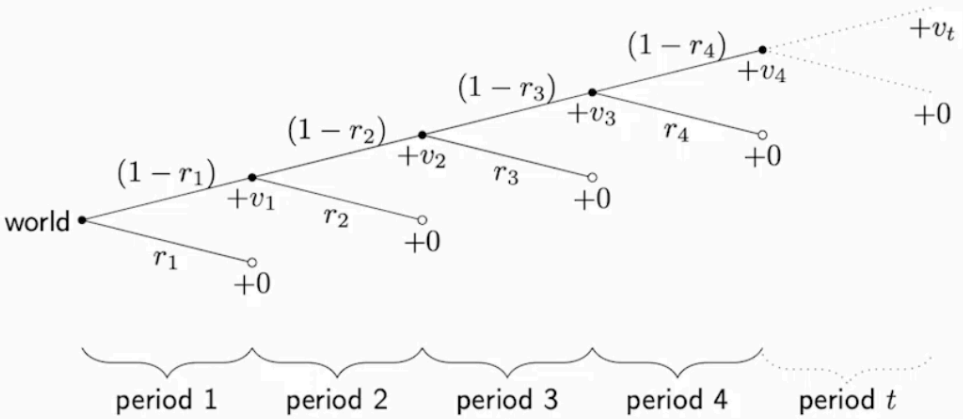
Those are gonna be the two drivers.
How big is the future, and then how much value is in that future?

And thinking about these two, sort of, things that impact the value of the future, you can get some kind of unintuitive results.

So as Toby Ord and David Thorstad have pointed out, that if the baseline level of existential risk is really high, the cost-effectiveness of x-risk mitigation goes down. Why is that? Because even if you avoid one particular catastrophic event, if there are a bunch of other ones waiting in the wings, you shouldn't expect the future to be that long, and therefore your payoff is gonna be less.

So these can interact in sometimes surprising ways. So here's how Arvo went about modeling the value of the future.

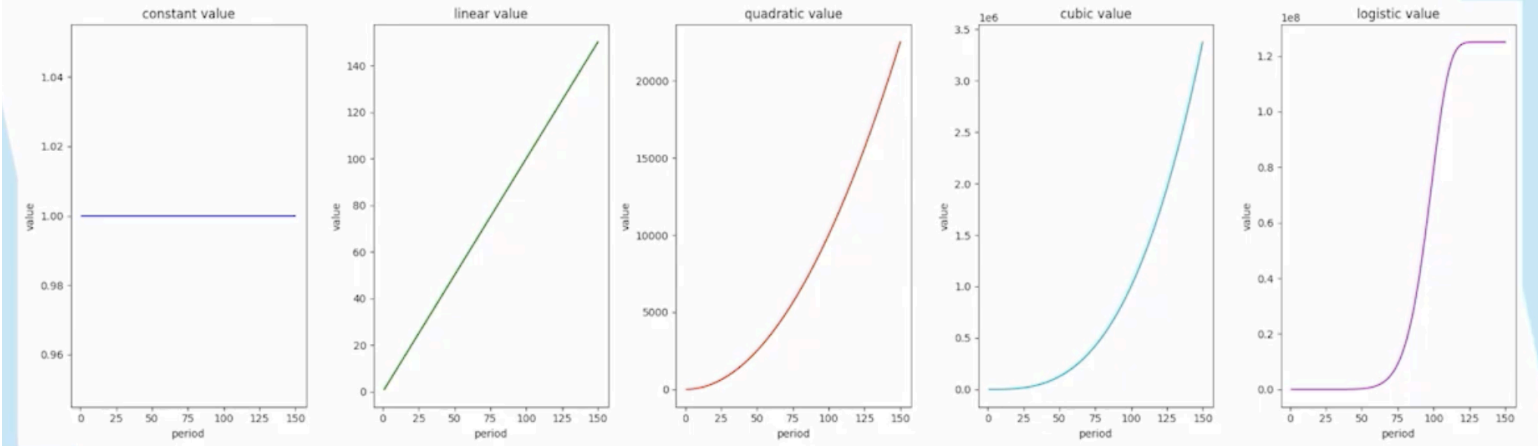
Modeling the expected value of the future



So take our world now, and then take some next time step.
Next period, it could be a year, it could be a hundred years, whatever. There's some chance, r1,
that there's an extinction-level event, in which case we get no more value in that time.
The amount of value in the world goes down to zero. But then with probability 1 - r, we make it to the next step, and we get a certain amount of value in that time, the value that accrues when humanity survives for, say, another hundred years.

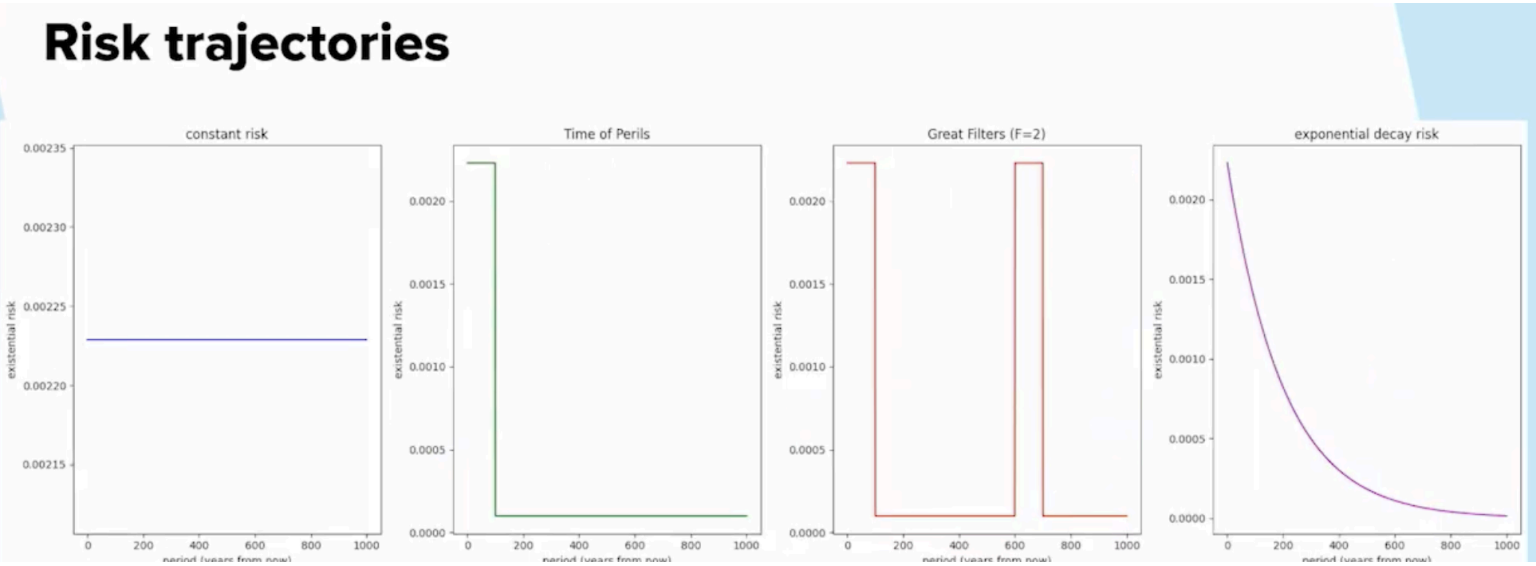
Then you can iterate that process. Then the next period, there's some chance that we go extinct, there's some chance that we continue to persist and get that amount of value. Using this sort of stepwise process,
we can sort of model, how much is the amount of risk gonna change over this time period as things evolve, and how much does the change in value per period of time also change?

Growth trajectories



So Arvo considered five growth trajectories, different assumptions you can make about how the amount of value per period will change over the future. On the constant growth model, I won't walk through all of these, but I'll give you some examples, we assume that basically the amount of growth that we've seen is gonna continue into the future, but at a sort of constant, sorry, the constant one just says we have the same amount of value per time period. We could also assume that value is gonna increase linearly. In the future, we're gonna get more value for [a given] time period. But you can also make more extreme assumptions, like cubic value. So you might think that in the future, we will create digital mines or we will colonize space. There'll just be so many more people that the amount of value that gets added per year in the future is much more than today.

He also considered various risk trajectories. How should we think of r as changing over time?



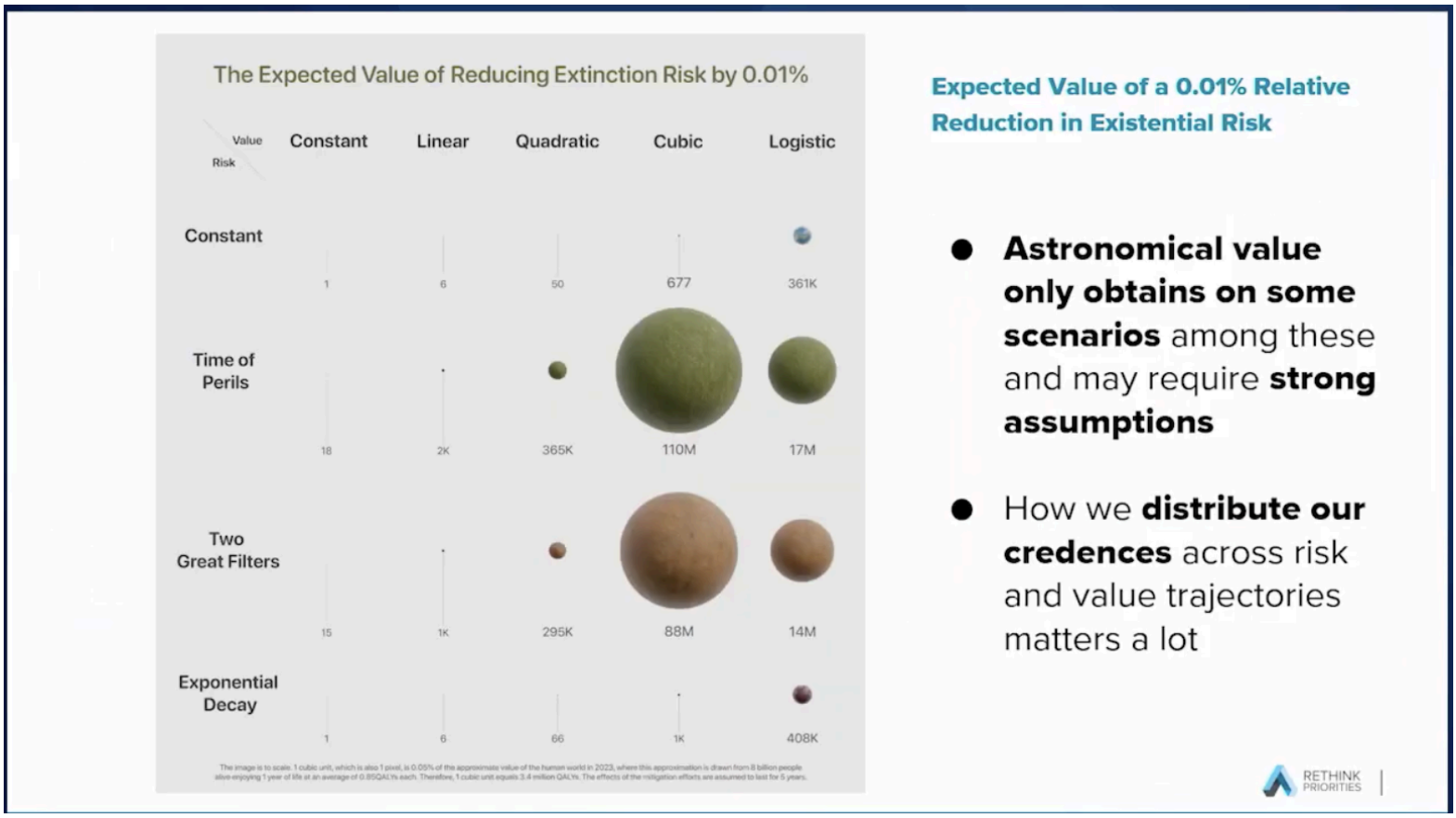
You might assume that there's constant risk, that there's always the same amount of existential risk that will be the same in 1,000 years as is today.

Another key assumptions you could make is what's called the time of perils, which says we are currently in a time of very high risk, but if we make out of this period, then the amount of risk is gonna go down significantly. So we're currently living in the precipice, but if we make it out of here, the risk in 1,000 years goes way down.

We can, sort of, generalize this and consider great filters kind of hypotheses that says humanity goes through periods of time with really high risk, and then after that, it settles down and goes through low risk, and humanity keeps kind of going through these cycles.

You also might think that if we colonize space, the amount of existential risk is gonna go down exponentially to the extent that we're distributed, and single events can't really take us out.

So Arvo combined all of these to model how different sets of assumptions affect our judgments about the amount of value exists in the future.



So he uses the constant-constant in the top left as sort of the baseline, everything's normalized to that. And then he says, how much value's in the future, based on other conjunctions of these assumptions? And notice that if you assume that we're living in a time of perils and there'll be cubic growth in the future, you get 110 million times more value in the future than you do if you assume a constant risk, constant growth set of assumptions.

So given how much the value of the future depends on these, like, really big assumptions and how little we know about these assumptions, it seems like you should be wary of saying, confidently, how much value exists in the future. Astronomical value only obtains on some scenarios, and they might require strong assumptions.

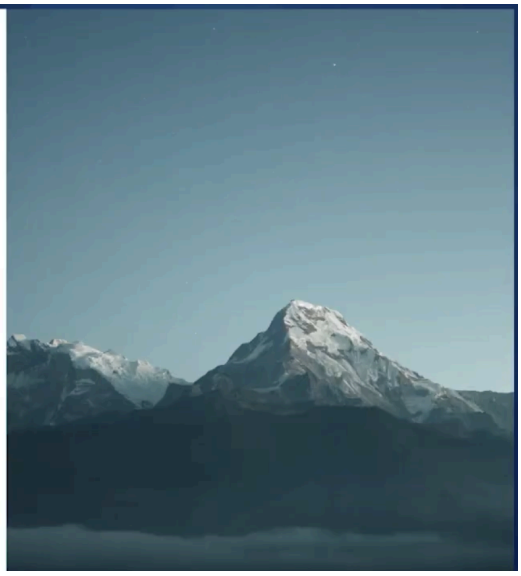
I won't go into this talk, I'll advertise, as part of the CURVE Sequence, David Bernard examined the time of perils hypothesis and whether or not we should think that it's probable.

How we distribute our credences across risk and value trajectories matters a lot, and I think a lot more research needs to be done here to say, how should we distribute these credences?

So what are the lessons that we can draw from this for cause prioritization in GCR?

Lesson 1: Trajectory Change

- Trajectory change might be more important than mere survival
- Not explored: value per individual, not just growth in # of individuals



One is that when you talk about the value of the future, that can change by millions and millions of times, depending on what that future is like.

This suggests that trajectory change might be more important for cost-effectiveness estimates than mere survival. It matters kind of a lot more how we persist in the future than whether we persist in the future

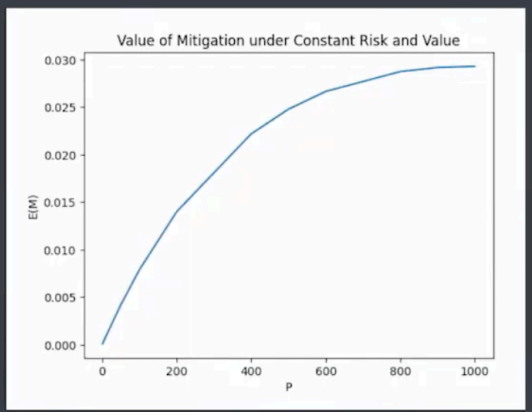
So if you can take actions that ensure that we're more likely to get in a cubic growth scenario or that the amount of risk goes down, you can yield a huge amount of value, whereas if you just ensure that we exist in a constant-constant kind of future, the value's gonna be a lot less.

When we considered the growth of the future, we just considered number of individuals. We didn't consider value per individual, but that would also probably make a huge difference to the value of the future.

A second thing that Arvo found out is that persistence was an extremely strong driver of expected value estimates.

Lesson 2: Persistence is Key

- Persistence: for how long does action lower the risk?
- Higher persistence → more value
- Diminishing marginal returns



So persistence is a measure of, for how many of those periods in his model did you lower the risk.

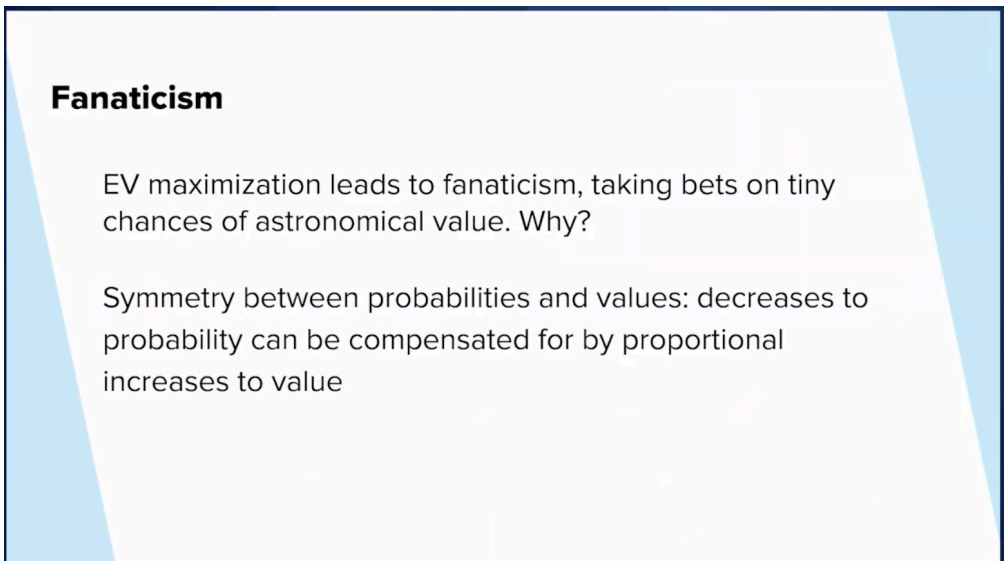
And we found that the actions that had higher persistence usually had much higher amounts of value, although they were diminishing marginal returns.

So for example, if you pass an action that's gonna be overturned by Congress in five years, such that the persistence of your effort was only five years, that had a much lower expected value than those that had effects on risk for, say, 100 years or more.

Those of you who really like expected value maximization can tune out for a bit.

So another main part of the project, we looked at alternatives to EV maximization to say what they might say about cross-prioritization in this area. So why might you go looking for alternatives?

One reason is that people are leery about EV maximization leading to fanaticism.



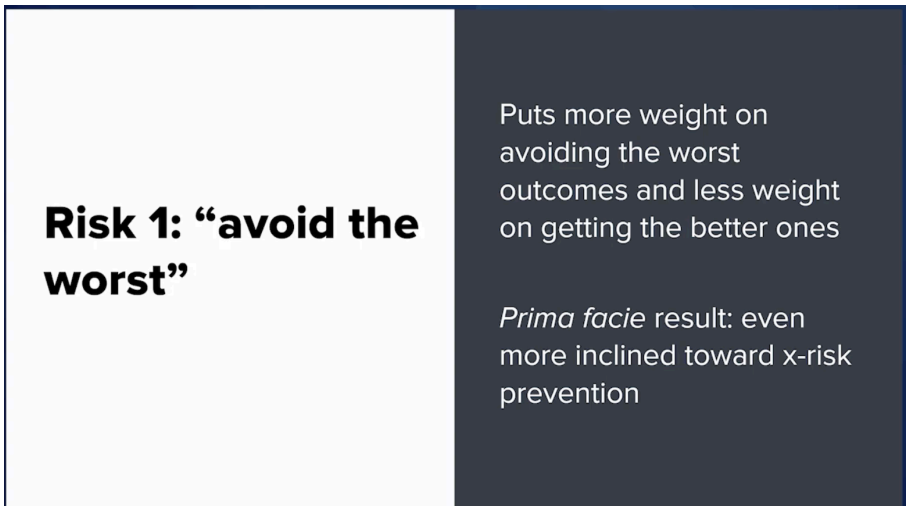
It recommends taking bets on tiny chances of astronomical value, as we've seen.

Why does it do this? Well, it posits a symmetry between probabilities and values.

So if you have the probability of a good effect but double the value of that good effect, that has the same expected value as your original action. Keep doing that and doing that, and as long as you have an astronomical value outcome, it's gonna recommend that you take any chance of getting that astronomical value, no matter how small.

Some people find that problematic and leading to Pascal's mugging case and stuff like that.
But you don't have to go that far. Most people are risk-averse in some sense and don't actually behave in accordance with EV maximization. So we might wanna ask, given that risk aversion is prominent, how might we model that, and what might you do if you have those attitudes? So to examine the effect of risk aversion, we found that there were really three kinds of risk aversion that were at play and were relevant to assessing actions in this space.

So the first kind of risk is what we called avoid-the-worst risk aversion.



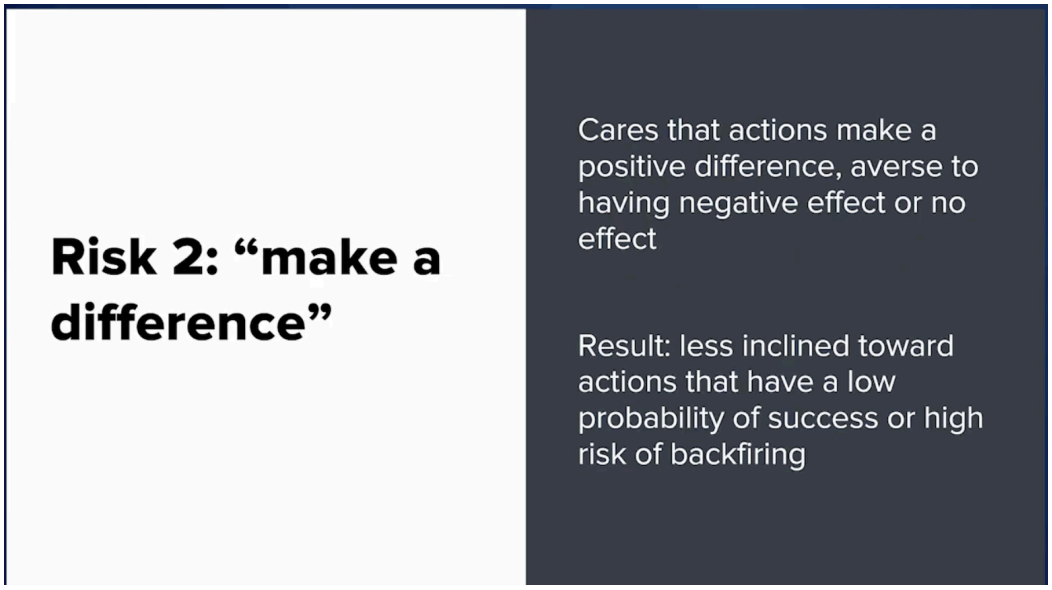
The most, sort of, standard philosophical and economic models are modeling this. If you're risk-averse in this way, you put more weight on avoiding the worst outcomes and less weight on getting the better ones. You're really concerned with avoiding the worst-case scenarios, and that's gonna, like, influence you more than the long shot of getting the best ones.

So if you ask how risk-averse are you, think about how early you got to the airport on the way here.

The prima facie result seems to be that you'll be even more inclined towards x-risk prevention if you're risk-averse in this way than you would be otherwise. We should play better safe than sorry. You'll be more influenced by the event that we all go extinct than the smaller gains of saving people from malaria, for example.

I'll problematize that in a bit, but it's good for now.

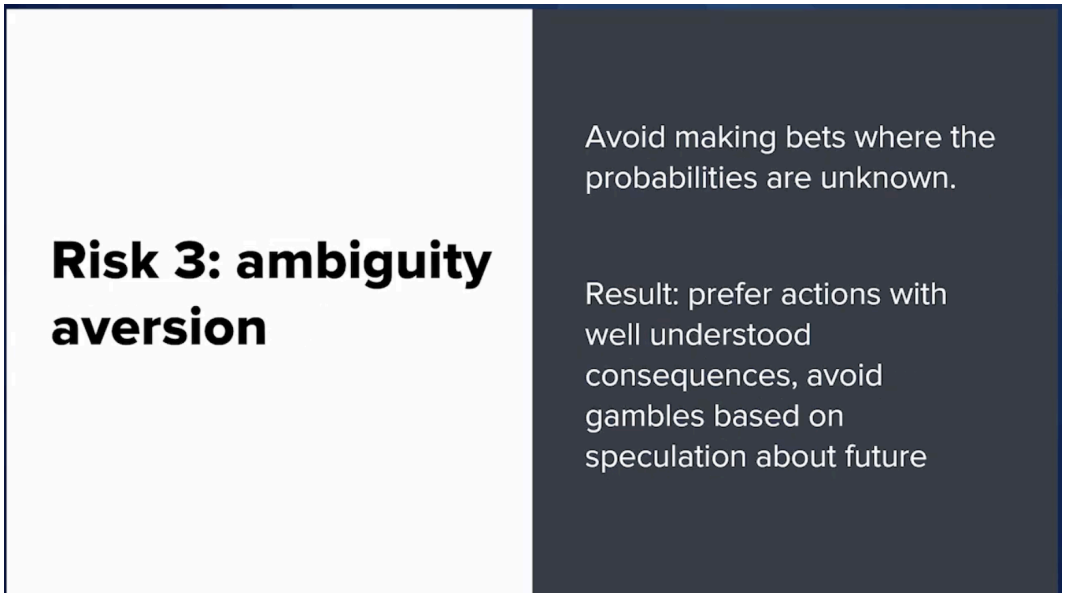
A second kind of risk aversion you might display is that you might be really concerned that your actions make a difference.



You might be really inclined to avoid situations where you've thrown your money away
and you've either made no difference or made things worse. That's the kind of risk aversion we also see.

So you're less focused on, like, what's the worst-case state of the world we can end in? But the worst case for you is that your money doesn't do anything. As a result, you'd be less inclined towards action
with a low probability of success or with a high probability of backfiring.

The last kind of risk aversion that we studied and modeled is ambiguity aversion or an aversion to uncertainty.



If you're risk-averse in this way, you don't like taking bets when you don't know the probabilities involved. You don't like taking bets when there's a lot of uncertainty about what the outcomes are and how probable they are.

As a result, ambiguity-averse individuals prefer actions with well understood consequences and avoid gambles based on just kind of ignorance and speculation about the future. So I and Laura Duffy both wrote a series of reports looking at formal models of all of these kinds of risk aversion and some of their implications for different causes.

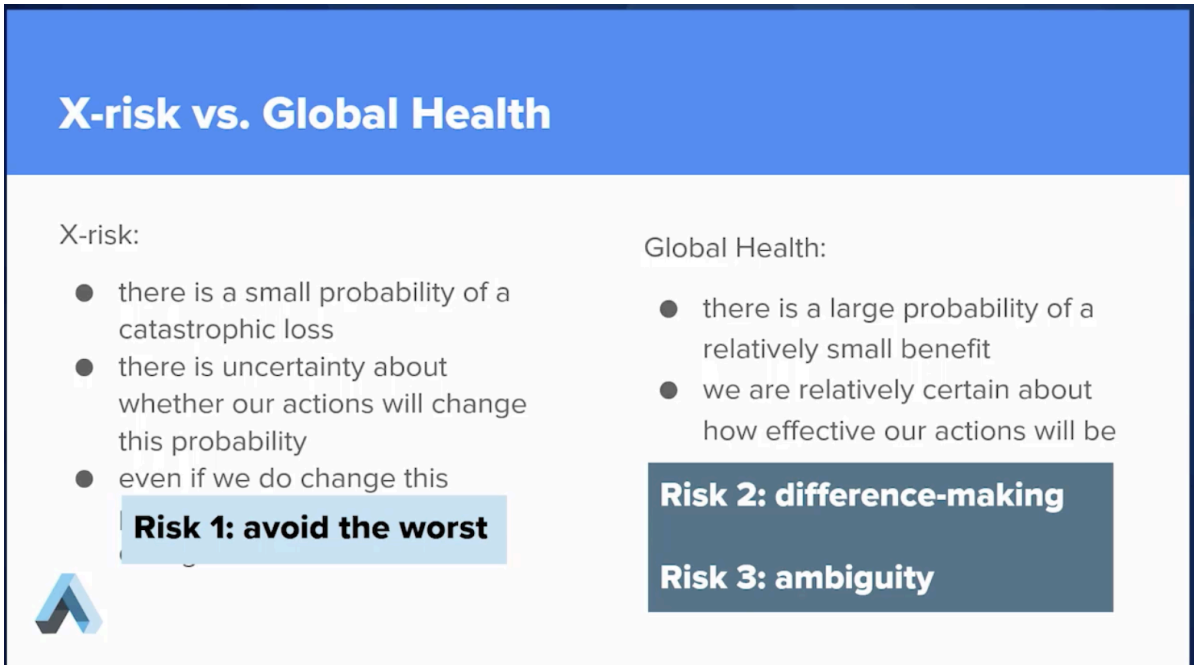
You can take a look at the details, but I'll give some sort of, like, qualitative results, sort of some general upshots.

So let's compare x-risk action at a high level versus global health actions at a high level.

With x-risk, there's a small probability, hopefully, of catastrophic loss. There's uncertainty about whether or not our actions will change this probability. And even if we do change this probability, we'll probably only change it by a little bit.

Compare that to global health where there's a large probability of a relatively small benefit and we're relatively certain, given how much research we've already done, about how effective our actions will be.

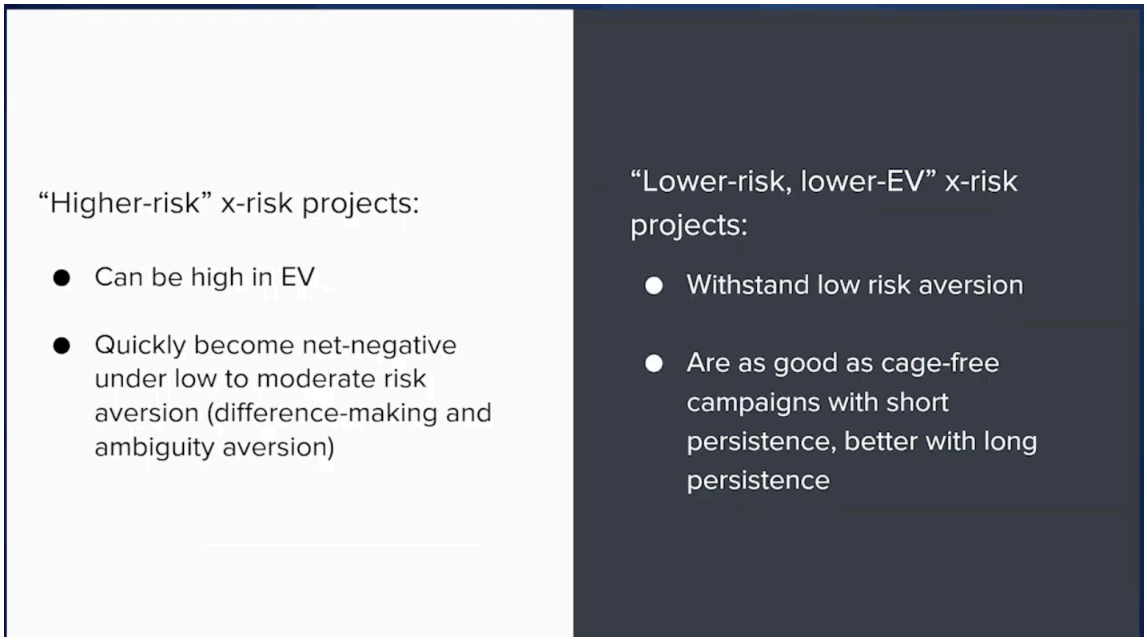
Interestingly, x-risk, well, taking these attitudes causes you to diverge even more so than EV maximization.



If you wanna avoid the worst, you'll probably be more inclined towards x-risk than you would've otherwise.

And if you were difference-making or ambiguity risk-averse, you'll be more likely to favor global health 'cause you like known probabilities and you like making a difference.

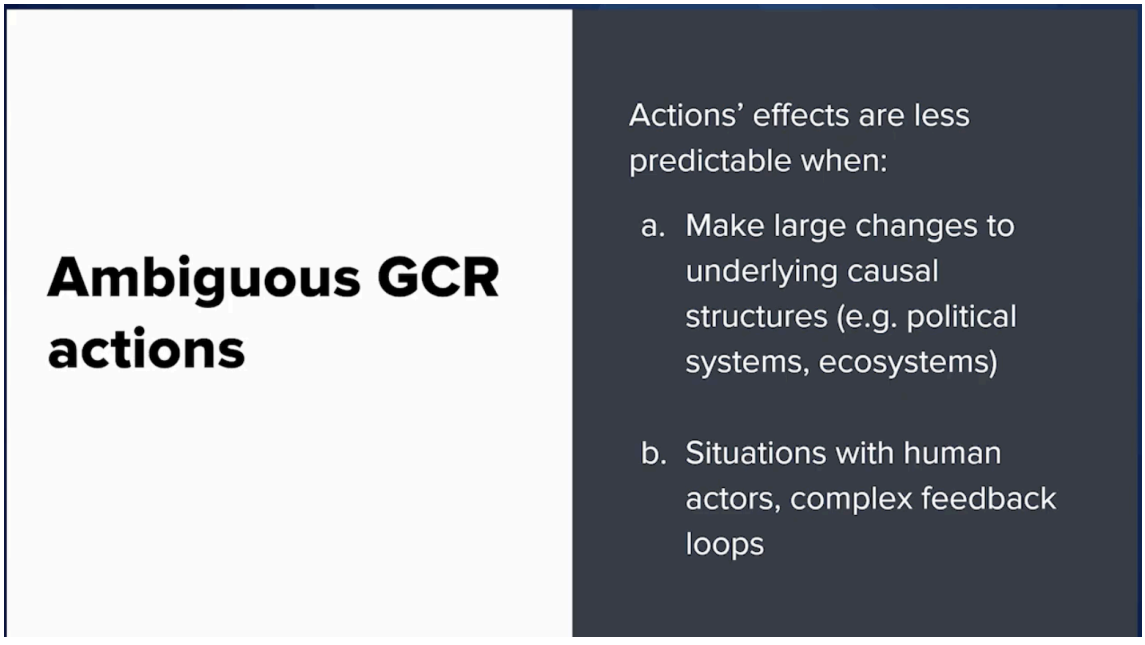
Within the GCR space, Laura did some good modeling about various higher risk x projects, much more speculative, have a higher risk of backfiring versus lower risk and potentially lower EV x-risk projects.



So higher risk projects quickly become net negative if we introduce either difference-making or ambiguity types of risk aversion. And even on some kinds of avoid-the-worst risk aversion, they look bad if the chance of backfiring is high enough. So risk aversion can really push us away from really speculative high-risk ones, even if they have a high expected value.

On the other hand, there are certain lower risk ones that do very well under risk aversion and even start to look as good as cage-free campaigns if they have long persistence.

Within the GCR space, if you're ambiguity-averse, you wanna take actions where you know the probabilities, where you have a good handle on what could happen.



Actions' effects are less predictable to the extent that they make large changes to underlying causal structures. So for example, if I take a hospital and give them PPE, I can kind of guess what's gonna happen. If I change democracy in the United States, who knows. I have a much better grip on, sort of, modular changes to known causal structures than I do to actions that result in upheavals of those causal structures, like ecosystem engineering, geoengineering, stuff like that.

And in particular, actions that involve human actors that have complex feedback loops also tend to be more ambiguous. So passing some sort of, like, nuclear war prevention act might make that more probable, given all the weird feedback loops of human actors.

So to sort of summarize how risk attitudes affect GCR cause prioritization, if you're an avoid-the-worst risk-averse person, you're even more inclined to be worried about extinction. It seems like you should be very focused on x-risk avoidance and risk-averse people have [been very focused on x-risk avoidance].

Two things that I wanna sort of, caveats I wanna raise.

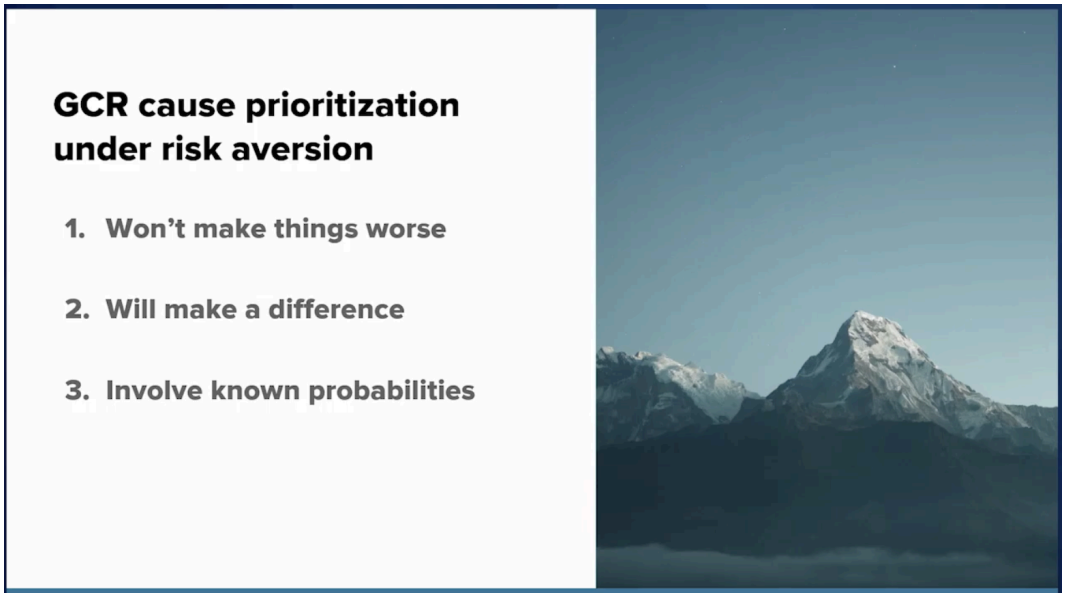
- The first is that you need to pay a lot of attention to the probability of backfire if you're risk-averse, even more so than if you're just an EV maximizer. Because if you make the worst-case scenario even more probable, then the risk-averse person's gonna really not be inclined to take that.
- Secondly, the worst-case scenario might not be where we go extinct. The worst-case scenario might be where we persist for a long time in a really, really, really bad state. And so it's a mistake to think that extinction should be equated to the worst-case scenario.

We should be focused on bad long-run futures, for example, where factory farming is even more prevalent with a growing population, where we mistreat digital mines in huge numbers, where we persist in, like, an AI dictatorship under which people are really miserable. Again, if you're risk-averse, maybe you should be more focused

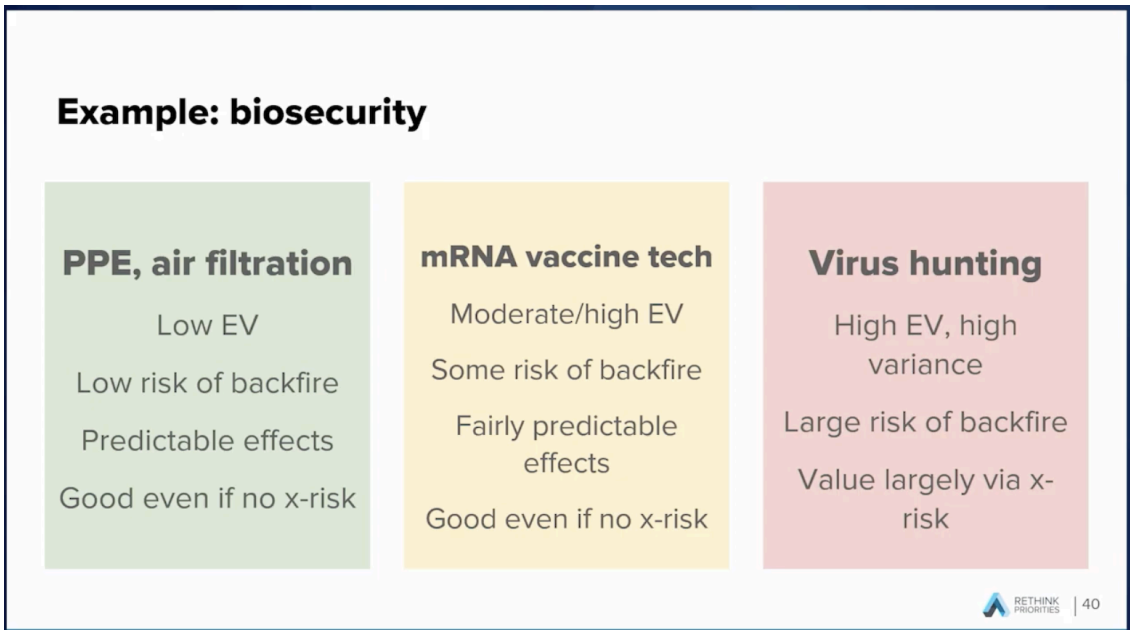
on how we persist rather than whether we persist. If you wanna make a difference, if that's what you're really worried about is cases where you don't have any effect or where you make things worse, you might wanna favor GCR outcomes that are going to have a high probability of having good outcomes, even if there is no existential risk sort of waiting in the wings. I'll give you some examples in the next slide.

And lastly, if you don't like making bets on uncertainty, you're gonna have this kind of attitude.
Sometimes x-risk actions have a high EV because we don't know what's gonna happen. So we give some small credence to some huge astronomical outcome. But if you're like, that potential value is arising from ignorance about what might happen, that we reserve some probability for some, you know, really, really good outcome, you're not gonna be swayed by that. What you want is, like, a firm probability that you actually have a chance of bringing about something really good. This is also gonna favor doing more research. You're gonna pay money to resolve your ambiguities about things and become more certain.

And so that seems to be something you should be inclined to do. Here's where things start to get a little speculative, so get out your salt shakers. Do any GCR actions farewell under all kinds of risk aversion? These are gonna be things, if they exist, that won't make things worse, that will make a difference, and involve known probabilities.



So ranking things from, like, green looks good from the light of all kinds of risk aversion, and red looks kind of bad. Let's compare some actions in the biosecurity space.



[green, yellow, red]

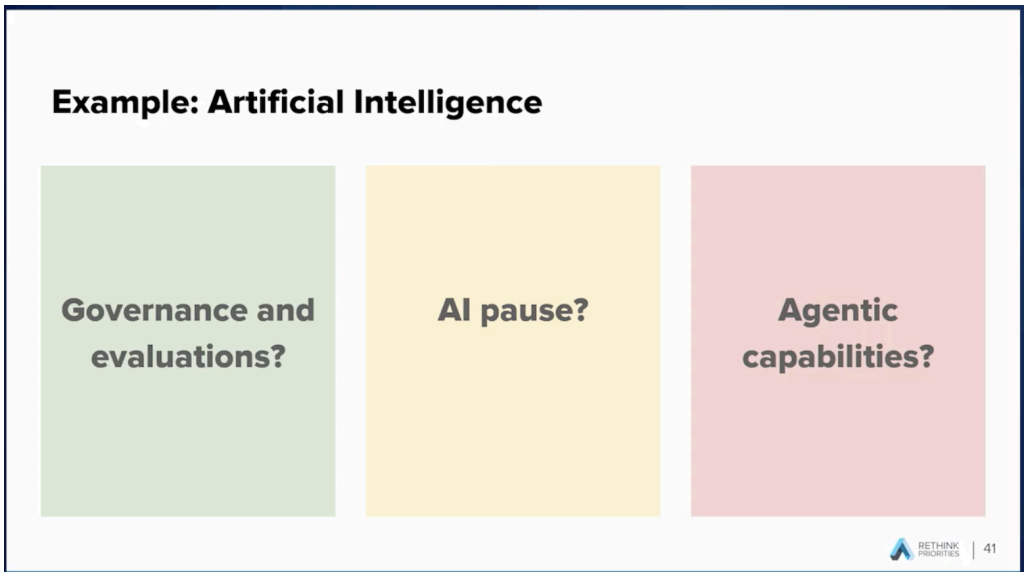
One kind of action that seems pretty good if you're risk-averse is something like personal protective equipment or air filtration. That has maybe a fairly low expected value. The potential upsides are not as enormous as for other actions, but has a very low risk of backfire. Maybe I have a limited imagination, but it's hard to see how that's gonna make a extinction-level pandemic more probable. It has fairly predictable effects, and it has good effects, even if there is no extinction-level event in the offing. Even if the future proceeds as expected, it's gonna be pretty good.

Something in the middle might be something like technology to rapidly deploy new mRNA vaccines in response to new viruses that emerge. That's gonna have pretty moderate to high EV. I mean, they just saved tens of millions of people during COVID, so that looks pretty good. It might have some risk of backfire if those technologies make it more easy to engineer viruses as well. We might be worried that's gonna have some downstream bad effects, but so far, those effects have been fairly predictable,
and again, they're good, even if we're not worried about extinction. That could save a lot of lives
in the global health space as well.

On the other hand, something like virus hunting or virus engineering, having labs try to make viruses that could produce the next extinction-level pandemic or looking around for those viruses, that could potentially have huge EV if we find these viruses before they occur and already have vaccines ahead of time. On the other hand, if that information falls into the wrong hands, that can be a really, really, really scary outcome.

We're gonna get even more speculative.

If you're in the artificial intelligence sphere, which I'm guessing a lot of you are, it'd be helpful to think through, how would you categorize certain kind of actions?



[green, yellow, red]

Governance and evaluations looks pretty risk-safe.
An AI pause might be a little bit riskier if pausing allows nefarious actors to get ahead of us.
And trying to design new age agentic capabilities and implement them, potentially high EV, and also is gonna look pretty bad by a lot of other kinds of risk aversion.

There's a lot there. I threw up, like, dozens of variables that you might think are relevant to this.
Different kinds of attitudes from making cause prioritization decisions, how do you keep these all all together?

I will recommend, if you wanna play with some of these things, our team, including our illustrious MC,
built a model, the cross-cause cost-effectiveness model, that you can go to and play around with
that allows you to explore the effect of changing all of these assumptions on cost-effective estimates. So it gives you a wide array of possible actions you could take. You can sort of give your own settings for your estimates of how much risk is reduced, how much value, stuff like that. And it'll allow you to make cost-effectiveness comparisons between different causes.

One thing that we hope people take away from this tool is being able to understand the effects of uncertainty, how changing some, you know, variables, some assumptions where we're not really sure what the truth is, how changing those assumptions can have very different effects on which actions are more and less cost-effective.

So we hope it's a way of kind of structuring ignorance, structuring uncertainty and starting conversations about, what are the cruxes that we need to examine a little bit more to really make these decisions?

Thank you.

(audience applauding)

Thank you, Hayley. So as a reminder, we are now going to be moving into the questions part of this session, and you can submit questions through Swapcard.
You can also, if you look through the submitted questions, you can upvote questions that you would most like to hear the answers to.

So first question, any chance you can share these slides or a summary set of links about the work this is based on?

Yes, I have no idea what venue would be most appropriate
for doing that, but I'll talk to the conference organizers about that.
A second thing is this is gonna be taped, so I'm sure you wanna hear my voice blaring all this at you as well. And the third is that the CURVE Sequence, where you can see much more in-depth reports exploring all of this, is posted on RP's [Rethink Priorities'] website, and also we had as a sequence on the EA forum. If you just look for the CURVE Sequence, it's all compiled there. So all of those reports can be found there, including some that I didn't talk about because they were less squarely on the question of cause prioritization in GCR.

Is it important to consider the external perceptual and community-building value of having various cause areas and interventions in the EA portfolio? For example, many EA community members became interested in EA first through the lens of global health and poverty reduction then, over time, became interested in other areas, such as x-risk.

That's a fantastic question. It's something that our team has been thinking a lot about,
about, you know, you might have a set of decision theoretical principles
that leads you to a certain conclusion. How much should you also be sensitive towards the long-term future of EA given the kind of causes that we choose, should that ever recommend choosing causes that are outside what your tools recommend?

A few things to say about this. One is one of the reasons that we wanted to include risk aversion
is that a lot of people with NEA are really committed to expected value maximization, even when it leads to what look like on intuitive conclusions. But a large part of the general populace is risk-averse
and finds these risk attitudes to be extremely intuitive and a potential way out of some of these unintuitive consequences of, you know, EV-till-I-die kind of views.

So what we wanted in part by investigating risk was to say, hey, if you're within EA
and you really want, like, good formal tools, you want things to be quantifiable, you wanna be able to track, you know, certain intuitions rather than just going on vibes alone, there are other decision theoretical models that you could use that give you some of that formal precision and rigor that might give you some alternatives to EVM.

So one reason also is about the sort of perception. To the extent that you can explore attitudes
that are more in line with, you know, the general populace, you might wanna do so.

A second thing to say is that's just a really hard philosophical question. If you're like, hey, I've committed to EVM and I'm committed to say, like, existential risk work is the only important thing to be doing right now, if that's really the dyed-in-the-wool kind of person you are, then there's this interesting question of, how much are you able to benefit what you think is valuable in the long-run by being more pluralistic in the short-term?

And the last thing is, a lot of our work sort of undermine this idea that, like, even if you aren't EV maximizer, x-risk work is the only thing you should be doing. I think that should spur conversations about how much, EAs might be sort of, like, ignoring animal welfare and certain global health causes as also being extremely valuable, even by the lights of these, like, hardcore value maximization tools.

How much risk aversion are you building into the risk-averse analysis?
Are they based on empirical amount of risk aversion that individuals have in behavioral studies?
Or if not, how much risk aversion did you build in?

Good, so the short answer is we allow you to vary that. So we didn't use just one setting of risk aversion.

In particular, Laura Duffy has this really long report that you can go read where she walks through all of this in a lot of depth, it's really great work, where she says, like, for low, medium and high amounts of risk aversion with respect to each of these three kinds, what would the results be?

So the precise answer is gonna depend on the risk model that you use. I won't go super into it.
We looked at two of the best risk models we have, risk weighted expected utility theory and weighted linear utility theory, and then we gave various parameter settings for risk aversion, and you can see how those play out.

If I'm not mistaken, the CCM model that you can play with also allows you
to adjust how much risk aversion.

It has a few different settings.

Yeah. But the question of, like, how much risk aversion is a realistic setting, and how does that translate into the parameters of the model, can't really do service that question here, but recommend go especially look at Laura Duffy's how does risk aversion affects cause prioritization report.

Did working on this project change your opinions about risk aversion?

Yeah, it made it more complicated, yeah. So I take myself to be a fairly, like, worst-case scenario risk-averse person. I find I'm quite inclined towards that. I went through a brief period where I thought that make-a-difference risk aversion was really a plausible kind of decision strategy, that we should evaluate actions by the probability that they have a good effect rather than no or a bad effect. I'm less inclined towards difference-making risk aversion these days. I'll have a report published
on the EA forum soon about that.

I think the thing that I really came out was how much more ambiguity-averse I was at the end of this
than I was at the start. So again, I realize that sort of, you know, if I'm like, hey, you know, x-risk work is gonna have astronomical value, I had to ask myself, that amount of value is coming from the very long tail of the possibilities. That value is coming from the one-in-a-million chance that I make a difference.
How confident I am that that probability is one in a million versus one in a hundred thousand versus one in a billion, it really, sort of, quantified, like, a lot of these value estimates are depending on very precise settings of variables about which I have no information. So it made me more ambiguity-averse and more aware of, wow, more leery of making actions based on that kind of uncertainty.

As an individual donor or cause area contributor, what lessons should I take away from the CURVE report?

That's a good question. That's gonna partially depend on what your attitudes are. One result that came through loud and clear was animal welfare really looks good by the lights of everything we looked at. So again, I mean, there's this sort of, you know, false dichotomy we've been making between global health will help a few individuals a little bit with high certainty versus one that's going to have a low probability of affecting a huge number of individuals, and animals are sitting right in the middle of really large probability of affecting a really large number of individuals.

And so that's, like, on all of these decision theories, looking really, like, a great, not even compromised position, but looking really good. Also, on all kinds of risk aversion, helping animals really stuck out as an incredibly important thing. You can be very confident you're gonna have an effect. The worst-case scenario that currently exists, I mean, it's really bad, and we are seeing it right now. And we have good research on, like, not perfect, but good research on what makes animal lives go better and worse.

So the biggest thing that came out for me was let's get out of this framework of just comparing global health and x-risk work and think, like, animals are actually coming out, animal work is looking really good by the lights of all of these procedures. I hadn't anticipated just how strongly it would come out.

Do you think it's worth exploring the consequences of putting different amounts of weight into each type of risk aversion?

Yeah, it's interesting. So the risk models interact in interesting ways, and in particular, avoid-the-worst risk aversion and make-a-difference risk aversion very frequently come apart. So for example, in x-risk work, if you're really worried about the worst-case scenario, you're gonna be really inclined towards giving to low chances of, you know, saving us from catastrophe, where if you wanna make a difference, you really care about that, you're really gonna be averse to cases where you threw your money away at a small chance. You're gonna want more incremental gains.

In another study that we did, if you're really worried about avoiding the worst-risk aversion, you should be giving all your money to make the lives of shrimp better, even if you're not certain that they're sentient, because it'd be really bad if they were, whereas if you really wanna make a difference, give to humans. I know I'm making a difference if I make your lives better. So these kinds of risk aversion in important key cases, sometimes they overlap, but often they come apart, and so that's gonna be sort of a key pinch point, that those don't get combined interesting ways.

Adding ambiguity to the mix often does jive pretty well. Like, you can get conjunctions of those views that I think are interesting. But if you really are like, what kind of risk-averse agent am I, if you find it plausible, that's gonna be one key point of departure. Do you wanna make a difference with your money, or do you really care about the worst-case state of the world?

So one traditional approach to handling uncertainty uses discount rates. Did the CURVE report explore any connection between risk aversion uncertainty and discount rates?

To my knowledge, no. Discount rates weren't one of the main tools that we used in the CURVE report. I know it's something that our team talks about and weighed in on. I might be wrong, but I don't think any of those were actually used, sort of, in this.

A few different things though that are similar to discount rates or might kind of approximate it, one thing that's interesting you can play around with in the CM is, what if I only considered the, like, 99% of most probable outcomes? What's the expected value then, leaving off the really far ends of the tails? Or what's gonna happen if I only focus on the 95% of most probable cases? So that's not quite a discount rate, but it is a way of saying, like, what's the expected value if I ignore the really long tails of the distribution and just think about what's probably gonna happen?

And again, other things that you can do, so, you know, another report from the CURVE report talked a lot about, given that we don't know what's gonna happen in the far future, given that our information decays over time and we have a Bayesian prior that our actions have zero effect, what should we estimate the value of long-run actions are? So again, that's something that's wasn't really couched in terms of discount rates but approximates that sort of, like, what penalty should we give something when our knowledge decays over the future? But I'll say it's something our team is working on about proper discount rates for animal welfare cases and stuff like that. So it's in the air, it's not in the CURVE.

You looked at how different kinds of attitudes can favor various causes that are currently under consideration by effective altruists. Are there configurations of assumptions that you think would imply cause areas that are not currently focused on by EAs?

That's really interesting. So one of them that I mentioned here was x-risk work. Or, like, what's the probability we end up in a really, really, really bad future trajectory? That seems to be something that some EAs are talking about, of course, but it's not as common. To phrase things in terms of not just avoiding catastrophic risk but trying to increase the probability of a really good long-term future.

Dovetailing with what I said a few answers ago, thinking about animals in the long-run future is also a really important thing that has received some attention, but should receive some more. So if you think that the vast majority of value in this world is currently possessed by non-human creatures, and that will continue to be the case into the future - focusing on how much value exists in the future, not just limited to humans, I think, would be a really important cause area for people to work on.