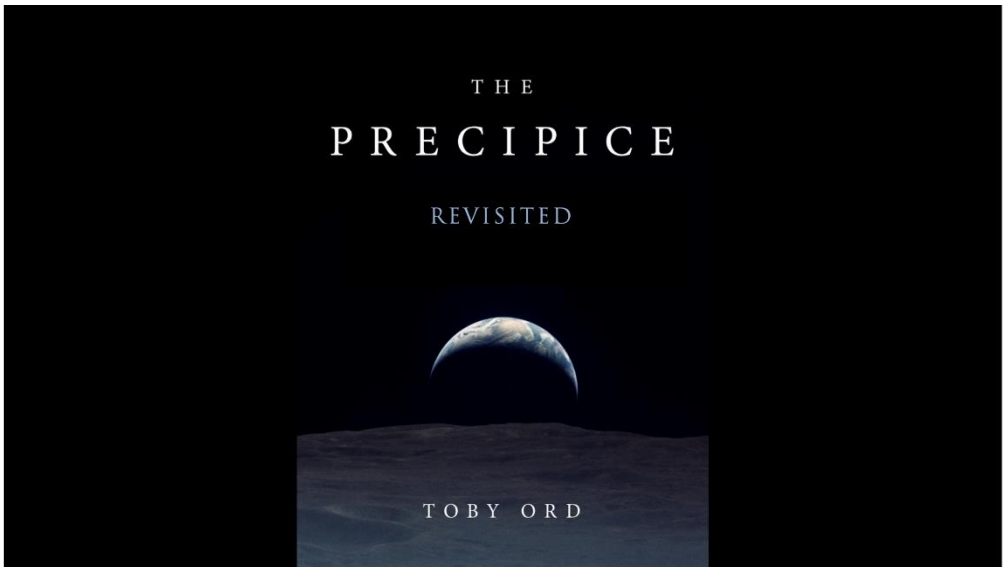


The Precipice Revisited https://www.tobyord.com/writing/the-precipice-revisited

In the years since I wrote *The Precipice*, the question I'm asked most is how the risks have changed. It's now almost four years since the book came out, but the text has to be locked down a long time earlier, so we are really coming up on about five years of changes to the risk landscape.

I'm going to dive into four of the biggest risks — climate change, nuclear, pandemics, and AI — to show how they've changed. Now a *lot* has happened over those years, and I don't want this to just be recapping the news in fast-forward. But luckily, for each of these risks I think there are some key insights and takeaways that one can distill from all that has happened. So I'm going to take you through them and tease out these key updates and why they matter.

I'm going to focus on changes to the landscape of *existential risk* — which includes human extinction and other ways that humanity's entire potential could be permanently lost. For most of these areas, there are many other serious risks and ongoing harms that have also changed, but I won't be able to get into those. The point of *this* talk is to really narrow in on the changes to *existential risk*.

CLIMATE CHANGE

Let's start with climate change.

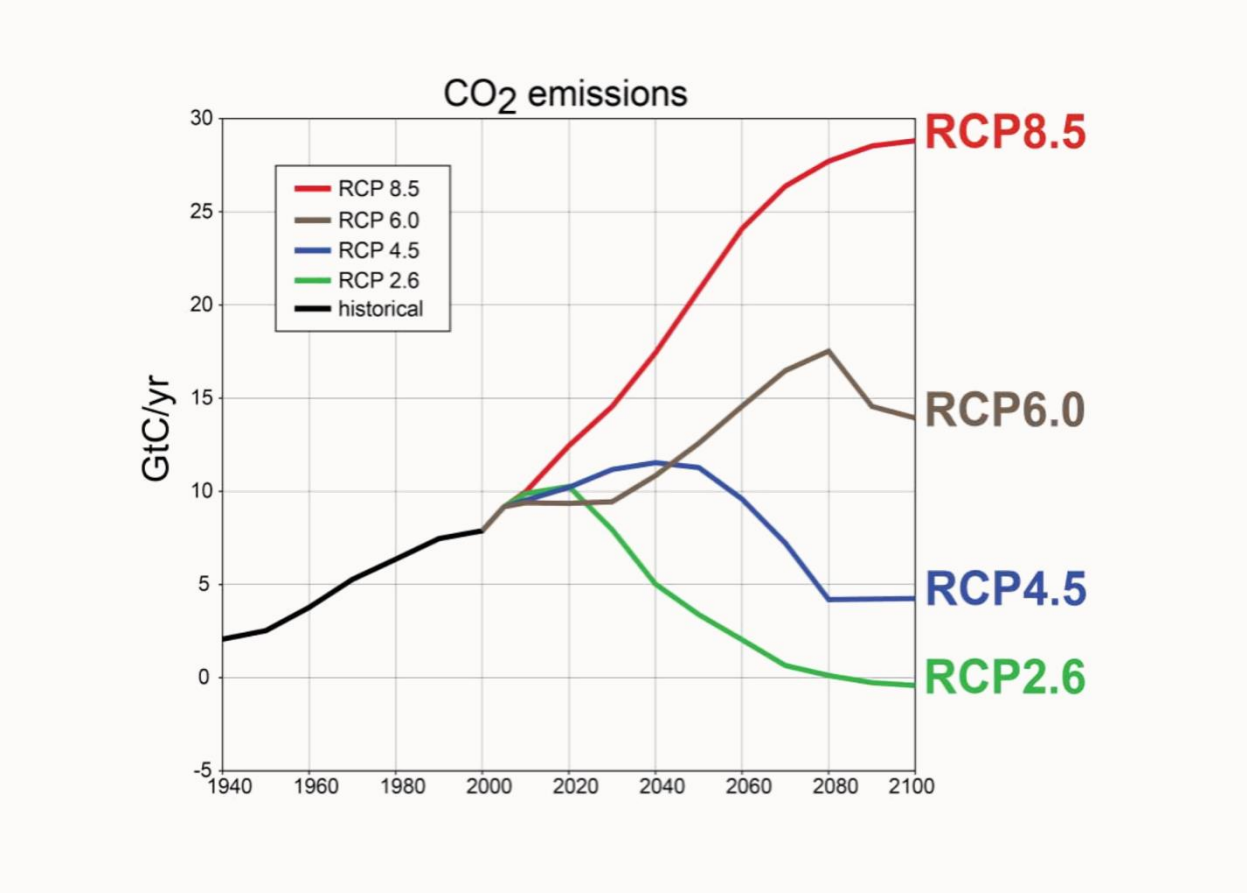
We can estimate the potential damages from climate change in three steps:

1. how much carbon will we emit?
2. how much warming does that carbon produce?
3. how much damage does that warming do?

And there are key updates on the first two of these, which have mostly flown under the radar for the general public.

Carbon Emissions

The question of how much carbon we will emit is often put in terms of Representative Concentration Pathways (RCPs).



Initially there were 4 of these, with higher numbers meaning more greenhouse effect in the year 2100. They are all somewhat arbitrarily chosen — meant to represent broad possibilities for how our emissions might unfold over the century. Our lack of knowledge about which path we would take was a huge source of uncertainty about how bad climate change would be. Many of the more dire climate predictions are based on the worst of these paths, RCP 8.5. It is now clear that we are not at all on RCP 8.5, and that our own path is headed somewhere between the lower two paths.

This isn't *great* news. Many people were hoping we could control our emissions faster than this. But for the purposes of *existential risk* from climate change, much of the risk comes from the worst case possibilities, so even just moving towards the middle of the range means lower existential risk — and the lower part of the middle is even better.

Climate Sensitivity

Now what about the second question of how much warming that carbon will produce? The key measure here is something called the *equilibrium climate sensitivity*. This is roughly defined as how many degrees of warming there would be if the concentrations of carbon in the atmosphere were to double from pre-industrial levels. If there

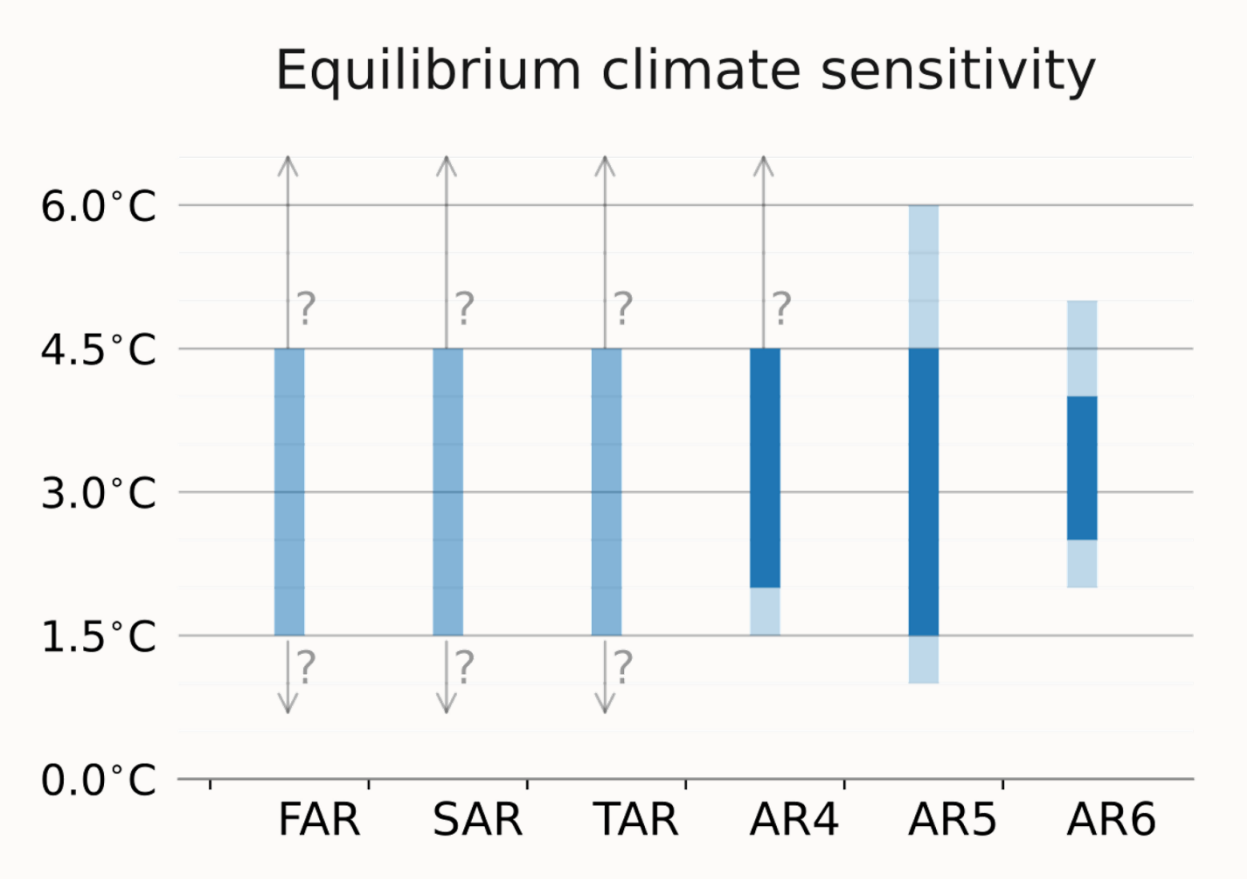
were no feedbacks, this would be easy to estimate: doubling carbon dioxide while keeping everything else fixed produces about 1.2°C of warming. But the climate sensitivity also accounts for many climate feedbacks, including water vapour and cloud formation. These make it higher and also much harder to estimate.

When I wrote *The Precipice*, the IPCC stated that climate sensitivity was likely to be somewhere between 1.5°C and 4.5°C.

When it comes to estimating the impacts of warming, this is a vast range, with the top giving three times as much warming as the bottom. Moreover, the true sensitivity could easily be even higher, as the IPCC was only saying that there was a two-thirds chance it falls within this range. So there was about a 1 in 6 chance that the climate sensitivity is even higher than 4.5°. And that is where a lot of the existential risk was coming from — scenarios where the climate is more responsive to emissions than we expect.

To understand the existential risk from climate change, we'd really want to find out where in this vast range the climate sensitivity really lies. But there has been markedly little progress in this. The range of 1.5° to 4.5° was first proposed the year I was born and had not appreciably changed in the following 40 years of research.

Well, now it has.



In the latest assessment report, the IPCC has narrowed the likely range to span from 2.5° to 4°, reflecting an improved understanding of the climate feedbacks. This is a mixed blessing, as it is saying that very good and very bad outcomes are both less likely to happen. But again, reducing the possibility of extreme warming will lower

existential risk, which is our focus here. And they also narrowed their 90% confidence range too, with them now saying it is 95% likely that the climate sensitivity is less than 5°.

In both cases here, the updates are about narrowing the range of plausible possibilities, which has the effect of reducing the chance of extreme outcomes, which in turn reduces the existential risk. But there is an interesting difference between them. When considering our emissions, humanity's actions in responding to the threat of climate change have lowered our emissions from what they would have been with business as usual. Our actions lowered the risk. In contrast, in the case of climate sensitivity, the change in risk is not causal, but evidential — we've discovered evidence that the temperature won't be as sensitive to our emissions as we'd feared. We will see this for other risks too. Sometimes we have made the risks lower (or higher), sometimes we've discovered that they always were.

NUCLEAR

Heightened Chance of Onset

At the time of writing, the idea of nuclear war had seemed like a forgotten menace — something that I had to remind people still existed. Here's what I said in *The Precipice* about the possibility of great-power war, when I was trying to make the case that it could still happen:

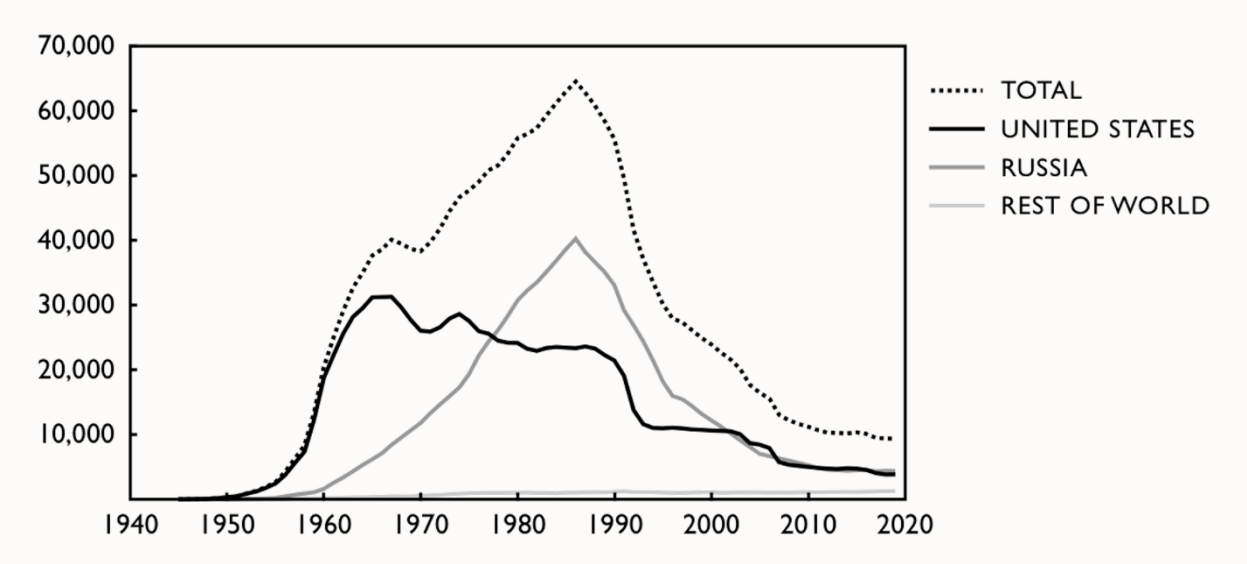
Consider the prospect of great-power war this century. That is, war between any of the world's most powerful countries or blocs. War on such a scale defined the first half of the twentieth century, and its looming threat defined much of the second half too. Even though international tension may again be growing, it seems almost unthinkable that any of the great powers will go to war with each other this decade, and unlikely for the foreseeable future. But a century is a long time, and there is certainly a risk that a great-power war will break out once more.

The Russian invasion of Ukraine has brought nuclear powers substantially closer to the brink of nuclear war. Remember all the calls for a no-fly zone, with the implicit threat that if it was violated, the US and Russia would immediately end up in a shooting war in Ukraine. Or the heightened rhetoric from both the US and Russia about devastating nuclear retaliation, and actions to increase their nuclear readiness levels. This isn't a Cuban missile crisis level, and the tension has subsided from its peak, but the situation is still markedly worse than anyone anticipated when I was writing *The Precipice.*

Likely New Arms Race

When I wrote *The Precipice*, the major nuclear treaty between the US and Russia (New START) was due to lapse and President Donald Trump intended to let that happen. It was re-signed at the last minute by the incoming Biden administration. But it can't be renewed again. It is due to expire exactly 2 years from now and a new replacement treaty would have to be negotiated at a time when the two countries have a terrible relationship and are actively engaged in a proxy war.

In *The Precipice*, I talked about the history of nuclear warhead stockpiles decreasing ever since 1986 and showed this familiar graph.



We'd been getting used to that curve of nuclear warheads going ever downwards, with the challenge being how to make it go down faster.

But now there is the looming possibility of stockpiles *increasing*. Indeed I think that is the most likely possibility. In some ways this could be an even more important time for the nuclear risk community, as their actions may make the difference between a dramatic rise and a mild rise, or a mild rise and stability. But it is a sad time.

In *The Precipice*, one reason I didn't rate the existential risk from nuclear war higher was that with our current arsenals there isn't a known scientific pathway for an all out war to lead to human extinction (nor a solid story about how it could permanently end civilisation). But now the arsenals are going to rise — perhaps a long way.

Funding Collapse

Meanwhile, there has been a collapse in funding for work on nuclear peace and arms reduction — work that (among other things) helped convince the US government to create the New START treaty in the first place. The MacArthur Foundation had been a major donor in this space, contributing about 45% of all funding. But it pulled out of funding this work, leading to a massive collapse in funding for the field, with nuclear experts having to leave to seek other careers, even while their help was needed more than at any time in the last three decades.

Each of these changes is a big deal.

- A Russian invasion of a European country, leading to a proxy war with the US.
- The only remaining arms control treaty due to expire, allowing a new nuclear arms race.
- While the civil society actors working to limit these risks see their funding halve.

Nuclear risk is up.

PANDEMICS

Covid

Since I wrote the book, we had the biggest pandemic in 100 years. It started in late 2019 when my text was already locked in and then hit the UK and US in full force just as the book was released in March.

Covid was not an existential risk. But it did have implications for such risks.

First, it exposed serious weaknesses in our ability to defend against such risks. One of the big challenges of existential risk is that it is speculative. I wrote in *The Precipice* that if there was a clear and present danger (like an asteroid visibly on a collision course with the Earth) things would move into action. Humanity would be able to unite behind a common enemy, pushing our national or factional squabbles into the background. Well, there was some of that, at first. But as the crisis dragged on, we saw it become increasingly polarised along existing national and political lines. So one lesson here is that the uniting around a common threat is quite a temporary phase — you get a few months of that and need to use them well before the polarisation kicks in.

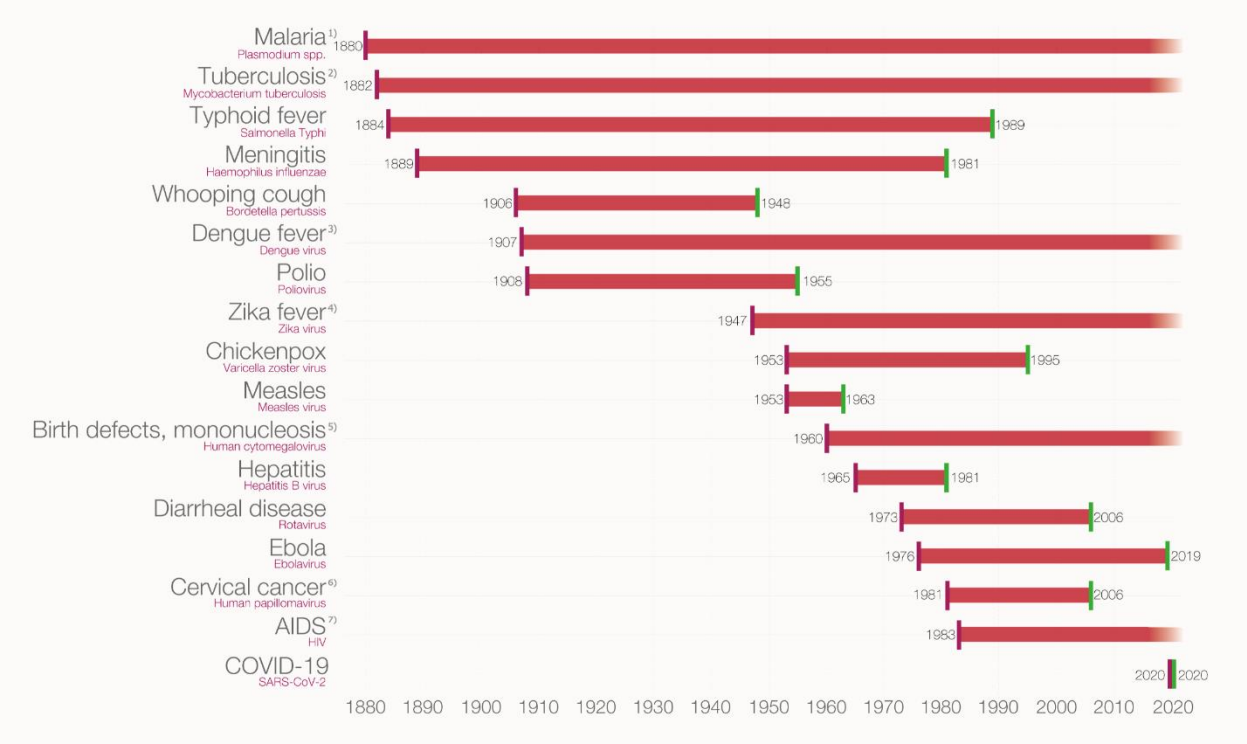
Another kind of weakness it exposed was in our institutions for managing these risks. I entered the pandemic holding the WHO and CDC in very high regard, which I really think they did not live up to. One hypothesis I've heard for this is that they did have the skills to manage such crises a generation ago, but decades without needing to prove these skills led them to atrophy.

A second implication of Covid is that the pandemic acted as a warning shot. The idea that humanity might experience large sweeping catastrophes that change our way of life was almost unthinkable in 2019, but is now much more beleivable. This is true both on a personal level and a civilisational level. We can now imagine going through a new event where our day to day life is completely upended for years on end. And we can imagine events that sweep the globe and threaten civilisation itself. Our air of invincibility has been punctured.

That said, it produced less of an immune response than expected. The conventional wisdom is that a crisis like this leads to a panic-neglect cycle, where we oversupply caution for a while, but can't keep it up. This was the expectation of many people in biosecurity, with the main strategy being about making sure the response wasn't too narrowly focused on a re-run of Covid, instead covering a wide range of possible pandemics, and that the funding was ring-fenced so that it couldn't be funnelled away to other issues when the memory of this tragedy began to fade.

But we didn't even see a panic stage: spending on biodefense for future pandemics was disappointingly weak in the UK and even worse in the US.

A third implication was the amazing improvement in vaccine development times, including the successful first deployment of MRNA vaccines.



When I wrote *The Precipice*, the typical time from vaccine development was decades and the fastest ever time was 10 years. The expert consensus was that it would take at least a couple of years for Covid, but instead we had several completely different vaccines ready within just a single year. It really wasn't clear beforehand that this was possible. Vaccines are still extremely important in global biodefense, providing an eventual way out for pandemics like Covid whose high R0 means they can't be suppressed by behaviour change alone. Other interventions buy us the time to get to the vaccines, and if we can bring that time sooner, it also makes that task of surviving that long easier.

A fourth and final implication concerns gain of function research. This kind of biological research where scientists make a more deadly or more transmissible version of an existing pathogen is one of the sources of extreme pandemics that I was most concerned about in *The Precipice*. While the scientists behind gain of function research are well-meaning, the track record of pathogens escaping their labs is much worse than people realise.

With Covid, we had the bizarre situation where the biggest global disaster since World War II was very plausibly caused by a lab escape from a well-meaning but risky line of research on bat coronaviruses. Credible investigations of Covid origins are about evenly split on the matter. It is entirely possible that the actions of people at one particular lab may have killed more than 25 million people across the globe. Yet the response is often one of indifference — that it doesn't really matter as it is in the past and we can't change it. I'm still genuinely surprised by this. I'd like to think that if it were someone at a US lab who may or may not be responsible for the deaths of millions of people that it would be the trial of the century. Clearly the story got caught up in a political culture war as well as the mistrust between two great powers, though I'd actually have thought that this would make it an even more visible and charged issue. But people don't seem to mind and gain of function research is carrying on, more or less as it did before.

Protective technologies

Moving on from Covid, there have been recent developments in two different protective technologies I'm really excited about.

The first is metagenomic sequencing. This is technology that can take a sample from a patient or the environment and sequence all DNA and RNA in it, matching them against databases of known organisms. An example of a practical deployment would be to use it for diagnostic conundrums — if a doctor can't work out why a patient is so sick, they can take a sample and send it to a central lab for metagenomic sequencing. If China had had this technology, they would have known in 2019 that there was a novel pathogen whose closest match was SARS, but which had substantial changes. Given China's history with SARS they would likely have acted quickly to stop it. There are also potential use-cases where a sample is taken from the environment, such as waste water or an airport air conditioner. Novel pandemics could potentially be detected even before symptoms manifest by the signature of exponential growth over time of a novel sequence.

The second is improving air quality. A big lesson from the pandemic is that epidemiology was underestimating the potential for airborne transmission. This opens the possibility for a revolution in public health. Just as there were tremendous gains in the 19th century when water-based transmission was understood and the importance of clean water was discovered, so there might be tremendous gains from clean air.

A big surprise in the pandemic was how little transmission occurred on planes, even when there was an infected person sitting near you. A lot of this was due to the impressive air conditioning and filtration modern planes have. We could upgrade our buildings to include such systems. And there have been substantial advances in germicidal UV light, including the new UVC band of light that may have a very good tradeoff in its ability to kill germs without damaging human skin or eyes.

There are a couple of different pathways to protection using these technologies. We could upgrade our shared buildings such as offices and schools to use these all the time. This would limit transmission of viruses even in the first days before we've become aware of the outbreak, and have substantial immediate benefits in preventing spread of colds and flu. And during a pandemic, we could scale up the deployment to many more locations.

Some key features of these technologies are that:

- They can deal with a wide range of pathogens including novel ones (whether natural or engineered). They even help with stealth pathogens before they produce symptoms.
- They require little prosocial behaviour-change from the public.
- They have big spillover benefits for helping fight everyday infectious diseases like the flu.
- And research on them isn't dual use (they systematically help defence rather than offence).

AI in Biotech

Finally, there is the role of AI on pandemic risk. Biology is probably the science that has benefitted most from AI so far. For example, when I wrote *The Precipice*, biologists had determined the structure of about 50 thousand different proteins. Then DeepMind's AlphaFold applied AI to solve the protein-folding problem and has used it to find and publish the structures of 200 million proteins. This example doesn't have any direct effect on existential risk, but a lot of biology does come with risks, and rapidly accelerating biological capabilities may open up new risks at a faster rate, giving us less time to anticipate and respond to them.

A very recent avenue where AI might contribute to pandemic risk is via language models (like ChatGPT) making it easier for bad actors to learn how to engineer and deploy novel pathogens to create catastrophic pandemics. This isn't a qualitative change in the risk of engineered pandemics, as there is already a strong process of democratisation of biotechnology, with the latest breakthroughs and techniques being rapidly replicable by people with fewer skills and more basic labs. Ultimately, very few people want to deploy catastrophic bioweapons, but as the pool of people who *could* expands, the chance it contains someone with such motivations grows. AI language models are pushing this democratisation further faster, and unless these abilities are removed from future AI systems, they will indeed increase the risk from bioterrorism.

AI

Which brings us, at last, to AI risk.

So much has happened in AI over the last five years. It is hard to even begin to summarise it. But I want to zoom in on three things that I think are especially important.

RL agents ⇒ language models

First, there is the shift from reinforcement learning agents to language models.

Five years ago, the leading work on artificial general intelligence was game-playing agents trained with deep reinforcement learning ('deep RL'). Systems like AlphaGo, AlphaZero, AlphaStar, and OpenAI 5. These are systems that are trained to take actions in a virtual world which outmanoeuvre an adversary — pushing inexorably towards victory no matter how the opponent tries to stop them. And they all created AI agents which outplayed the best humans.

Now, the cutting edge is generative AI: systems that generate images, videos, and (especially) text.

These new systems are not (inherently) agents. So the classical threat scenario of Yudkowsky & Bostrom (the one I focused on in *The Precipice*) doesn't directly apply. That's a big deal.

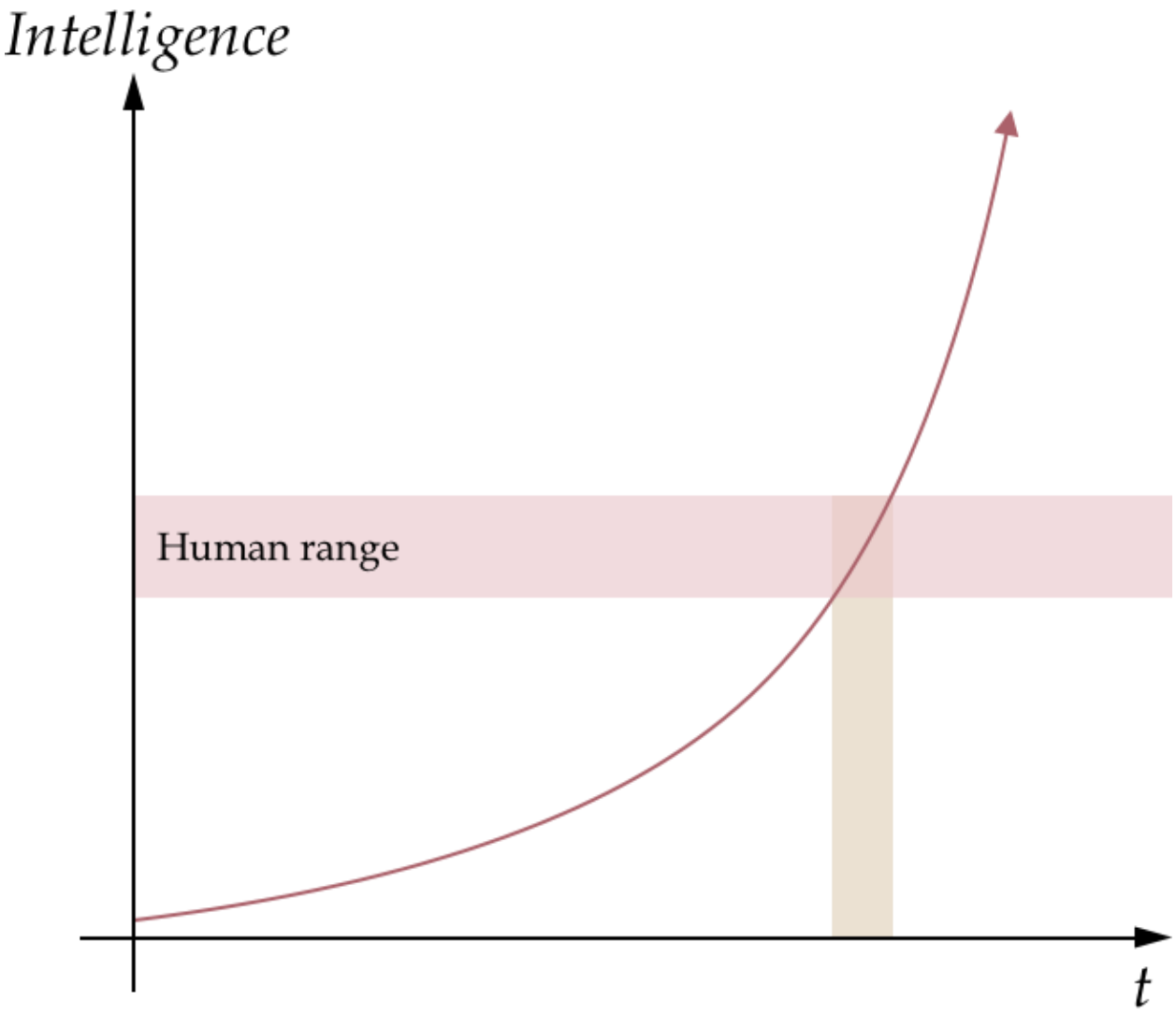
It does look like people will be able to *make* powerful agents out of language models. But they don't *have to* be in agent form, so it may be possible for first labs to make aligned non-agent AGI to help with safety measures for AI agents or for national or international governance to outlaw advanced AI agents, while still benefiting from advanced non-agent systems

And as Stuart Russell predicted, these new AI systems have read vast amounts of human writing. Something like 1 trillion words of text. All of Wikipedia; novels and textbooks; vast parts of the internet. It has probably read more academic ethics than I have and almost all fiction is nothing but humans doing things to other humans combined with judgments of those actions.

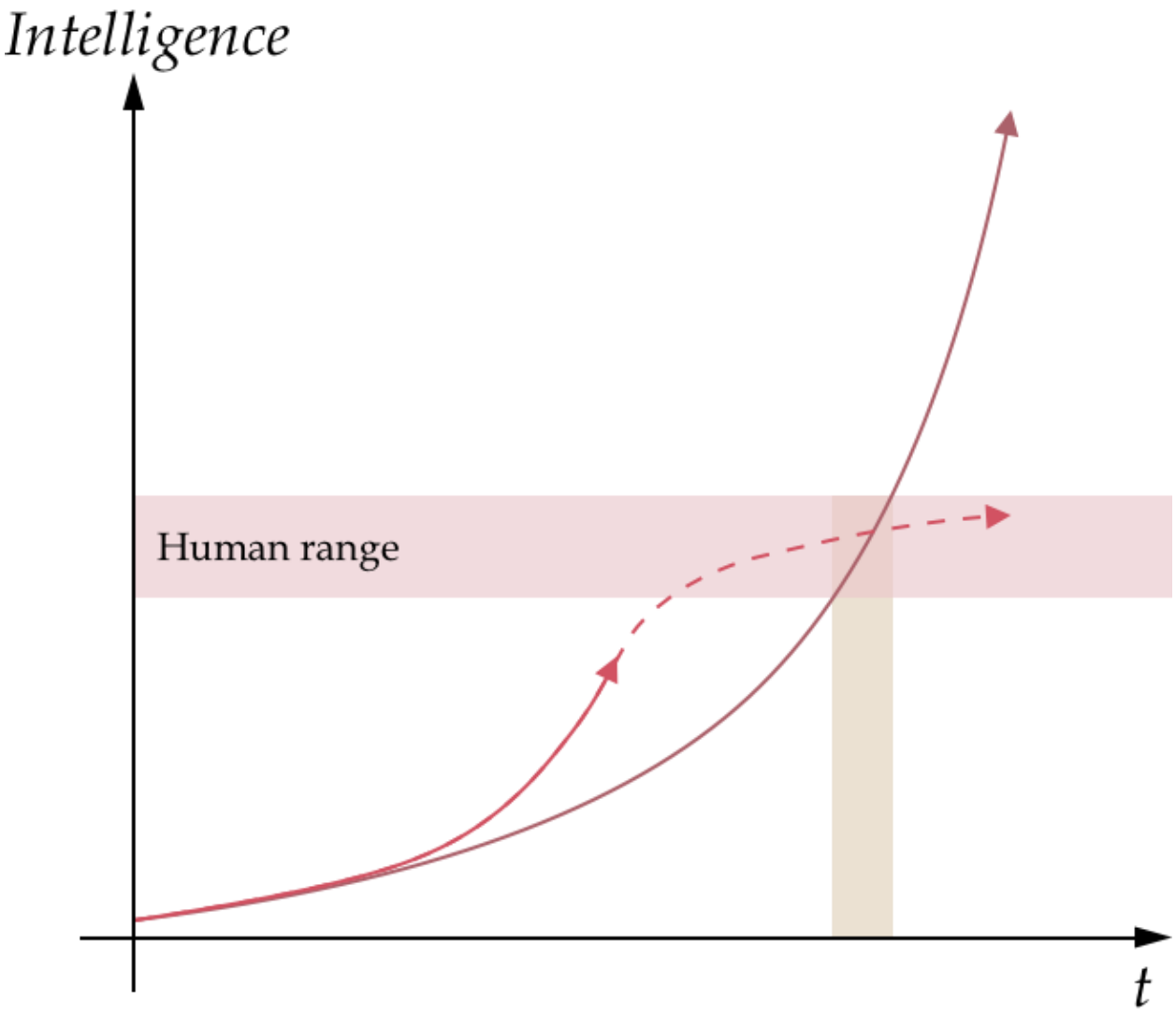
These systems thus have a vast amount of training signal about human values. This is in big contrast to the game-playing systems, which knew nothing of human values and where it was very unclear how to ever teach them about such values. So the challenge is no longer getting them enough information about ethics, but about making them *care*.

Because this was supervised learning with a vast amount of training data (not too far away from all text ever written), this allowed a very rapid improvement of capabilities. But this is not quite the general-purpose acceleration that people like Yudkowsky had been predicting.

He predicted that accelerating AI capabilities would blast through the relatively narrow range of human abilities because there is nothing special about those levels. And because it was rising so fast, it would pass through in a very short time period:

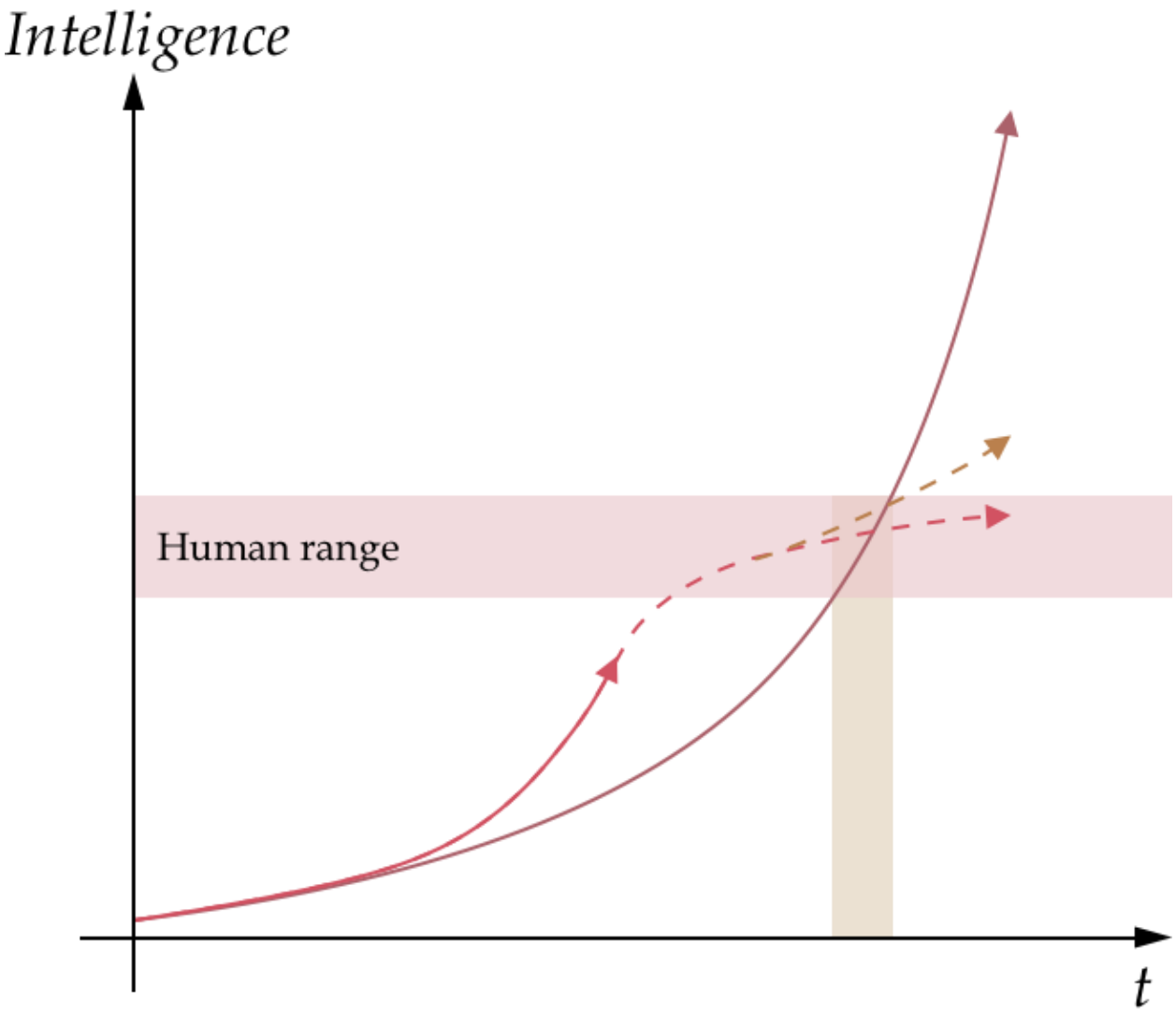


Language models have been growing more capable even faster. But with them there *is*something very special about the human range of abilities, because that is the level of all the text they are trained on. Their capabilities are being pulled rapidly upwards towards the human level as they train on all our writings, but they may also plateau at that point:



In some sense they could still exceed human capabilities (for example, they could know more topics simultaneously than any human and be faster at writing), but without some new approach, they might not exceed the best humans in each area.

Even if people then move to new kinds of training to keep advancing beyond the human level, this might return to the slower speeds of improvement — putting a kink in the learning curves:



And finally because language models rely on truly vast amounts of compute in order to scale, running an AGI lab is no longer possible as a nonprofit, or as a side project for a rich company. The once quite academic and cautious AGI labs have become dependent on vast amounts of compute from their corporate partners, and these corporate partners want rapid productisation and more control over the agenda in return.

Racing

Another big change in AI risk comes from the increased racing towards AI.

We had long been aware of the possibilities and dangers of racing: that it would lead to cutting corners in safety in an attempt to get to advanced AGI before anyone else.

The first few years after writing *The Precipice* saw substantial racing between OpenAI and DeepMind, with Anthropic also joining the fray. Then in 2023, something new happened. It became not just a race between these labs, but a race between the largest companies on Earth. Microsoft tried to capitalise on its prescient investments in OpenAI, seizing what might be the moment of its largest lead in AI to try for Google's crown jewel — *Search.*

They put out a powerful, but badly misaligned system that gave the world its first example of an AI system that turned on its own users.

Some of the most extreme examples were threatening revenge on a journalist for writing a negative story about it, and even threatening to kill an AI ethics researcher. It was a real *2001 A Space Odyssey* moment, with the AI able to autonomously find out if you had said negative things about it in other conversations and threaten to punish you for it.

The system still wasn't really an agent and wasn't really *trying* to get vengeance on people. It is more that it was simulating a persona who acts like that. But it was still systematically doing behaviours that if done by humans would be threats of vengeance.

Microsoft managed to tamp down these behaviours, and ended up with a system that did pose a real challenge for Google on search. Google responded by merging its Brain and DeepMind labs into a single large lab and to make them much more focused on building ever larger language models and immediately deploying them in products.

I'm really worried that this new race between big tech companies (who don't *get* the risks from advanced AI in the same way that the labs themselves do) will push the labs into riskier deployments, push back against helpful regulation, and make it much harder for the labs to pull the emergency brake if they think their systems are becoming too dangerous.

Governance

Finally, the last 16 months has seen dramatic changes in public and government interest.

By the time I wrote *The Precipice*, there was certainly public interest in the rise of AI — especially the AlphaGo victory over Le Sedol. And the US government was making noises about a possible race with China on AI, but without much action to show for it.

This really started to change in October 2022, when the US government introduced a raft of export controls on computer chip technologies to keep the most advanced AI chips out of China and delay China developing an AI chip industry of its own. This was a major sign of serious intent to control international access to the compute needed for advanced AI development.

Then a month later saw the release of ChatGPT, followed by Bing and GPT-4 in early 2023. Public access to such advanced systems led to a rapid rise in public awareness of how much AI had progressed, and public concern about its risks.

In May, Geoffrey Hinton and and Yoshua Bengio — two of the founders of deep learning — both came out expressing deep concern about what they had created and advocated for governments to respond to the existential risks from AI. Later that month, they joined a host of other eminent people in AI in signing a succinct public statement that addressing the risk of human extinction from AI should be a global priority.

This was swiftly followed by a similar statement by The Elders — a group of former world leaders including a former UN secretary general, WHO director-general, and presidents and prime ministers of Ireland, Norway, Mongolia, Mexico, Chile, and more with multiple Nobel Peace prizes between them. They agreed that: "without proper global regulation, the extraordinary rate of technological progress with AI poses an existential threat to humanity."

Existential risk from AI had finally gone from a speculative academic concern to a major new topic in global governance.

Then in October and November there was a flurry of activity by the US and UK governments. Biden signed an executive order aimed at the safety of the frontier AI systems. The UK convened the first of an ongoing series of international meetings on international AI governance, with an explicit focus on the existential risk from AI. 28 countries and the EU signed the Bletchley Declaration. Then the US and UK each launched their own national AI safety institutes.

All told, this was an amazing shift in the Overton window in just 16 months. While there is by no means universal agreement, the idea that existential risk from AI is a major global priority is now a respected standard position by many leaders in AI and global politics, as is the need for national and international regulation of frontier AI.

The effects on existential risk of all this are complex:

- The shift to language models has had quite mixed effects
- The racing is bad
- The improved governance is good

The overall picture is mixed.

CONCLUSIONS

In *The Precipice*, I gave my best guess probabilities for the existential risk over the next 100 years in each of these areas. These were 1 in 1,000 for Climate and for Nuclear, 1 in 30 for Pandemics and 1 in 10 for Unaligned AI. You are probably all wondering how all of this affects those numbers and what my new numbers are. But as you may have noticed, these are all rounded off (usually to the nearest factor of 10) and none of them have moved *that* far.

But I *can* say that in my view, Climate risk is down, Nuclear is up, and the story on Pandemics and AI is mixed — lots of changes, but no clear direction for the overall risk.

I want to close by saying a few words about how the world has been reacting to existential risk as a whole.

In *The Precipice*, I presented the idea of *existential security*. It is a hypothetical future state of the world where existential risk is low and kept low; where we have put out the current fires, and put in place the mechanisms to ensure fire no longer poses a substantive threat.

We are mainly in the putting-out-fires stage at the moment — dealing with the challenges that have been thrust upon us — and this is currently the most urgent challenge. But existential security also includes the work to establish the norms and institutions to make sure things never get as out of control as they are this century: establishing the moral seriousness of existential risk; establishing the international norms, then international treaties, and governance for tackling existential risk.

This isn't going to be quick, but over a timescale of decades there could be substantial progress. And making sure this process succeeds over the coming decades is crucial if the gains we make from fighting the current fires are to last.

I think things are moving in the right direction. The UN has started to explicitly focus on existential risks as a major global priority. The Elders have structured their agenda around fighting existential risks. One British Prime Minister quoted *The Precipice* in a speech to the UN and another had a whole major summit around existential risk from AI. The world is beginning to take security from existential risks seriously, and that is a very positive development.

Delivered as a talk: The Precipice Revisited, at EA Global: Bay Area, February 2024.