Case

The overwhelming consensus of AI researchers agrees that AGI is likely by 2047---prefer the largest survey with the most recency. Expert audits are better than specific warrants: we'll never be able to discuss every intricacy that makes AGI possible or not with limited time, but experts have spent years reviewing the evidence.

Grace '24 [Katja Grace, co-founder of AI Impacts, a research project trying to incrementally answer decision-relevant questions about the future of artificial intelligence (AI), .AI Impacts, Berkeley, CA; Harlan Stewart, AI Impacts; Julia Sandkühler, Psychology Professor at the University of Bonn; Stephen Thomas, AI Impacts; and Jan Brauner, Professor of Computer Science at Oxford; January 2024; "THOUSANDS OF AI AUTHORS ON THE FUTURE OF AI," doa: 11/20/24 https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf]

In the **largest survey** of its kind, we surveyed **2,778 researchers** who had **published** in **top-tier** artificial intelligence (AI) venues, asking for their predictions on the pace of AI progress and the nature and impacts of advanced AI systems. The aggregate forecasts give at least a 50% chance of AI systems achieving several milestones by 2028, including autonomously constructing a payment processing site from scratch, creating a song indistinguishable from a new song by a popular musician, and autonomously downloading and fine-tuning a large language model. If science continues undisrupted, the chance of unaided machines **outperforming** humans in **every possible task** was estimated at **10% by 2027**, and **50% by 2047**. The latter estimate is **13 years earlier** than that reached in a **similar survey** we conducted **only one year earlier** [Grace et al., 2022]. However, the chance of all human occupations becoming fully automatable was forecast to reach 10% by 2037, and 50% as late as 2116 (compared to 2164 in the 2022 survey).

AGI is coming, hard to control, and deadly for humanity.

Hoel '23 [Erik Hoel, their author, neuroscientist and assistant research professor at Tufts University, former postdoctoral researcher at Columbia University, Ph.D. in Neuroscience from the University of Wisconsin Madison; being interviewed by Russ Roberts from EconTalk; April 3, 2023; "Erik Hoel on the Threat to Humanity from AI," doa: 11/20/24 https://www.econtalk.org/erik-hoel-on-the-threat-to-humanity-from-ai/#audio-highlights] recut winky

Russ Roberts: I want to congratulate you. **You** are the first person who has actually caused me to be alarmed about the **implications of AI**--artificial intelligence--and the **potential threat** to **humanity**. Back in 2014, I interviewed Nicholas Bostrom about his book Superintelligence, where he argued AI could get so smart it could trick us into doing its bidding because it would understand us so well. I wrote a lengthy follow-up to that episode and we'll link to both the episode and the follow-up. So, I've been a skeptic. I've interviewed Gary Marcus who is a skeptic. I recently interviewed Kevin Kelly, who is not scared at all. But you--you--are scared. Last month you wrote a piece called "I Am Bing, and I Am Evil" on your Substack, The Intrinsic Perspective, and you actually scared me. I don't mean, 'Hmmm. Maybe I've underestimated the threat of AI.' It was more like I had a 'bad feeling in the pit of my stomach'-kind of scared. So, what is the central argument here? Why should we take this latest foray into AI, ChatGPT, which writes a pretty okay--a pretty impressive but not very exciting essay, can write some poetry, can write some song lyrics--why is it a threat to humanity? Erik Hoel: Well, I think to take that on very broadly, we have to realize where we are in the history of our entire civilization, which is that we are at the point where we are finally making things that are **arguably as intelligent** as a **human being**. Now, are they as intelligent right now? No, they're not. I don't think that these very advanced, large, language models that these companies are putting out could be said to be as intelligent as an expert human on whatever

subject they're discussing. And, the tests that we use to measure the progress of these systems supports that where they do quite well and quite surprisingly well on all sorts of questions like SAT [Standardized Achievement Test] questions and so on. But, one could **easily see** that **changing**. And, the big issue is around this concept of general intelligence. Of course, a chess-playing AI poses no threat because it's just slowly trained on playing chess. This is the notion of a narrow AI. Self-driving cars could never really pose a threat. All they do is drive cars. But, when you have a general intelligence, that means it's similar to a human in that we're good at all sorts of things. We can reason and understand the world at a general level. And, I think it's very arguable that right now, in terms of the generalness behind general intelligences, these things are actually more general than the vast majority of people. That's precisely why these companies are using them for search. So, we already have the general part quite well down. The issue is intelligence. These things hallucinate. They are not very reliable. They make up sources. They do all these things. And, I'm fully open about all their problems. Russ Roberts: Yeah. They're kind of like us, but okay. Yeah. Erik Hoel: Yeah, yeah, precisely. But, one could **easily imagine**, given the rapid progress that we've made just in the past couple years, that by **2025**, **2030**, you could have things that are both **more general** than a **human being** and as **intelligent** as any living person--perhaps **far more intelligent**. And, that enters this very scary territory, because we've **never existed** on the planet with **anything else like that**. Or, we did once a very long time ago, about 300,000 years ago. There's something like nine different species--or our cousins who we were related to--who were likely probably either as intelligent as us or quite close in intelligence. And they're all gone. And, it's probable that we exterminated them. And, then ever since then we have been the dominant masters and been no other things. And so, finally for the first time, we're at this point where we're creating these entities and we don't know quite **how smart** they can get. We simply have no notion. Human beings are very similar. We're all based on the same genetics. We might all be points stacked on top of one another in terms of intelligence and all the human beings and all the differences between people are all really just this zoomed-in minor differences. And, really you can have things that are vastly more intelligent. And if so, then we're at risk of either relegating ourselves to being inconsequential, because now we're living near things that are much more intelligent. Or alternatively, in the worst case scenarios, we simply don't fit into their picture of whatever they want to do. And, fundamentally, **intelligence** is the **most dangerous thing** in the universe. Atom bombs, which are so powerful, and so destructive and, in use of warfare so evil we've all agreed not to use them, are just this inconsequential downstream effect of being intelligent enough to build them. So, when you start talking about building things that are as or more intelligent than humans based on very different rules--things that are right now not reliable: they're unlike a human mind, we can't fundamentally understand them due to rules around complexity--and also, so far, they've **demonstrated empirically** that they can be **misaligned** and **uncontrollable**. So, unlike some people like Bostrom and so on, I think sometimes they will offer too specific of an argument for why you should be concerned. So, they'll say, 'Oh, well, imagine that there's some AI that's super-intelligent and you assign it to do a paperclip factory; and it wants to optimize the paperclip factory and the first thing it does is turn everyone into paperclips,' or something like that. And, the first thing when people hear these very **sci-fi arguments**, is to start **quibbling** over the particulars of like, 'Well, could that **really** happen?' and so on. But, I think the concern over this is this broad concern--that this is something we have to deal with, and it's going to be much like climate change or nuclear weapons. It's going to be with us for a very long time. We don't know if it's going to be a problem in five years. We don't know if it'll be a problem in 50 years. But it's going to be a problem at some point that we have to deal with. 7:17 Russ Roberts: So, if you're listening to this at home and you're thinking, 'It seems like a lot of doom and gloom, really it's too pessimistic'--I used to say things like, 'We'll just unplug it if it gets out of control,'--I just want to let readers know that this is a much better horror story than then Erik's been able to trace out in the first two, three minutes. Although I do want to say that, in terms of rhetoric, although I think there's a lot of really interesting arguments in the two essays that you wrote, when you talked about these other nine species of humanoids sitting around a campfire and inviting homo sapiens--that's us--into the circle and say, 'Hey, this guy could be useful to us. Let's bring him in. He could make us more productive. He's got better tools than we do,' that made the hair on the back of my neck stand up and it opened me to the potential that the other more analytical arguments might carry some water. Excuse me, carry some weight. So, one point you make, which is I think very relevant, is that all of this right now is mostly in the hands of profit-maximizing corporations who don't seem to be so worried about anything except novelty and cool and making money off it. Which is what they do. But, it is a little weird that we would just say, 'Well, they won't be evil, will they? They don't want to end humanity.' And you point out that that's really not something we want to rely on. Erik Hoel: Yeah. Absolutely. And, I think that this gets to the question of how should we treat this problem? And, I think the best analogy is to treat it something like climate change. And now, there is a huge range of opinion when it comes to climate change and all sorts of debate around it. But, I think that if you take the extreme end of the spectrum and say. 'There's absolutely no danger and there should be zero regulation around these subjects,' I actually think most people will disagree. They'll say, 'No, listen: this is something we do need to keep our energy usage as a civilization under control to a certain degree so we don't pollute streams that are near us,' and so on. And, even if you don't believe any specific model of exactly where the temperature is going to go--so maybe you think, 'Well, listen: there's only going to be a couple degrees of change. We'll probably be fine.' Okay? Or you might say, 'Well, there's definitely this doomsday scenario of a 10-degree change and it's so destabilizing,' and so on. Okay? But regardless, there are sort of reasonable proposals that one can do where we have to discuss it as a polity, as a group. You have to have an overarching discussion about this issue and make decisions regarding it. Right now with AI, there's no input from the public; there's no input from legislation; there's no input from anything. Like, massive companies are pouring billions of dollars to create intelligences that are fundamentally unlike us, and they're going to use it for profit. That's a description of exactly what's going on. Right now there's no red tape. There's no regulation. It just does not exist for this field. And, I think it's very reasonable to say that there should be some input from the rest of humanity when you go to build things that are as equally intelligent as a human. I do not think that that's unreasonable. I think it's something most people agree with--even if there are positive futures where we do build these things and everything works out and so on. Russ Roberts: Yeah. I want to--we'll come at the end toward what kind of regulatory response we might suggest. And, I would point out that climate change I think is a very interesting analogy. Many people think it'll be small enough that we can adapt. Other people think it is a existential threat to the future of life on earth, and that justifies everything. And, you have to be careful because there are people who want to get ahold of those levers. So, I want to put that to the side though, because I think you have more--we're done with that. Great--interesting--observation, but there's so much more to say. 11:35 Russ Roberts: Now, you got started--and this is utterly fascinating to me--you got started in your anxiety about this, and it's why your piece is called "I Am Bing, and I Am Evil," because Microsoft put out a chatbot, which is--I think internally goes by the name of Sydney--is ChatGPT-4, meaning the next generation pass what people have been using in the OpenAI version. And it was--let's start by saying it was erratic. You called it, earlier, 'hallucinatory.' That's not what I found troubling about it. I don't think it's exactly what you found troubling about it. Talk about the nature of what's erratic about it. What happened to the New York Times reporter who was dealing with it? Erik Hoel: Yes, I think a significant issue is that the vast majority of minds that you can make are completely insane. Right? Evolution had to work really hard to find sane minds. Most minds are insane. Sydney is obviously quite crazy. In fact, that statement, 'I Am Bing, and I Am Evil,' is not something I made up: It's something she said. This chatbot said, right? Russ Roberts: I thought it was a joke. I really did. Erik Hoel: Yeah. Yeah, no. It's something that this chatbot said. Now, of course, these are large, language models. So, the way that they operate is that they receive an initial prompt and then they sort of do the best that they can to auto-complete that prompt. Russ Roberts: Explain that, Erik, for people who haven't--I mentioned in the Kevin Kelly episode that there's a very nice essay by Steven Wolfram on how this might work in practice. But, give us a little of the details. Erik Hoel: Yeah. So, in general, the thing to keep in mind is that these are trained to auto-complete text. So, they're basically big artificial neural networks that guess at what the next part of text might be. And, sometimes people will sort of dismiss their capabilities because they think, 'Well, this is just like the auto-complete on your phone,' or something. 'We really don't need to worry about it.' But you don't--it's not that you need to worry about the text completion. You need to worry about the huge, trillion-parameter brain, which is this artificial neural network that has been trained to do the auto-completion. Because, fundamentally, we don't know how they work. Neural networks are mathematically black boxes. We have no fundamental insights as to what they can do, what they're capable of, and so on. We just know that this thing is very good at auto-completing because we trained it to do so. And, there's also no fundamental limit of what it can or can't learn. Like, for example, to auto-complete a story, you have to have a good understanding of human motivations. So, that means that this neural network that is trained on auto-complete now needs to understand things like human motivations in order to do auto-complete well. And, there are some analogies here. For example, there's a big subset of computational neuroscience, including the most-cited neuroscientist living--whose name is Karl Friston--who view the brain and argue that the brain is all based around minimizing the surprise of its inputs. Which is a very simple thing and looks a lot like auto-complete. So, I don't think that you can look at these things and say it's just auto-complete. It's not the auto-complete that's the problem. It's the huge neural network that's doing the auto-complete that could possibly be dangerous or at least do things that we don't expect, which is exactly what you're talking about with what happened with the release of Sydney, where there was all sorts of reports coming out of the crazy things that they were able to get this model to do, and say, and play-act as. Russ Roberts: Just to be clear on this auto-complete thing, which--that phrase makes it sound particularly unassuming about what it's capable of doing. You can correct me if I'm wrong. The way I understand it is: I might ask ChatGPT to write me a poem about love in the style of Dr. Seuss. So, it might start a sentence then with 'Love,' and then the next word that usually comes after love in human expression is 'is.' 'Love is a'--and now it's going to look at the millions and millions of sentences in its database called 'love is a,' and it's going to find, not necessarily--this is the coolest part about it--not necessarily the most common word that follows, because that would end up being after a while kind of flat. But, sometimes the most common and sometimes a surprise word, which gives us the feeling that it's actually doing something thoughtful. So, it might say, 'Love is a game,' or it might say, 'Love is a form of war'; or, it's going to look around and then it's going to keep going, and then it gets to an end, it's going to find, okay, after that sentence, what kind of sentence might come next, or what word would come next as the first word, etc. And, it's a slightly--just slightly--smarter, more effective version of my gmail--that, when I get a gmail at the bottom, it gives me three choices: 'Thanks'; 'Thanks so much'; 'I'd rather not.' And, in that sense, Gmail is smart. Not very smart, not very thoughtful. I usually don't take what it says, but sometimes I do when it's useful. 17:06 Russ Roberts: The real issue to me--one of the issues--and we're going to come back and talk about Sydney because we didn't really go into the erratic thing, because it's one of the creepiest things I've ever read. The autocomplete function is something like what we do as human beings. You could argue that's how we compose. Beethoven, in terms of musical composition, he always knew what note should come next and in a way, that's all ChatGPT does. But, that's all we do maybe when we write. We don't really understand--our brain is also a bit of a black box. So, I don't think we should then jump to the similarly--just because all it does is autocomplete doesn't mean it's not smart. But, also, I don't think we should say, because the brain also does a lot of effective auto completion we should assume it's a brain. It doesn't seem sentient. And, I'm curious--I know you talk about that in your second essay. So, if I'm the skeptic and I say, 'Well, okay, so it has this ability to pass an SAT test,' because it has a lot of data. And, I don't quite understand how, because it's a black box and it's a neural network and I can't model it cleanly. But, it's not sentient. It's not going to have desires. Erik Hoel: Before we move on to the question of sentience, because I think that that's a really deep, deep well, I just want to clarify a couple things about the actual operations of these systems. So, in terms of a metaphorical understanding of what's going on, the sort of thing, like, there's a big lookup table of the net probability of next words is a conceptual sort of description of what it's doing; but there is actually no lookup tables of the probabilities. What's actually happening is that there's this huge neural network, which are things designed based off of principles of how our own brains operate. Now, there's all sorts of differences; but the fundamentals of artificial neural networks--of what are called the artificial neural networks--were always based off of our real biological neural networks. So, there is this huge digital brain. It looks in structure, very different from our brain, but it's still based off of that. And, now we train this neural network to auto-complete text. So, that's what it does; but we don't know how it does it. We don't know where the probabilities of these words, sort of, are within the network. And, the way that we train it, people think that we're--a big misunderstanding is that people think that we are programming in responses or putting in information. And we're really not. And, I think a good analogy for how this is actually working would be: Imagine that there were alien neuroscientists who are incredibly more advanced than we are, and they want to teach a human being how to do math. So, they take some young kid and they put a math test in front of the young kid, and they have the young kid do the math test. And the kid gets 50% of the questions wrong. And, then the aliens, rather than trying to explain math to the student--the way that we would teach them--they just say, 'Okay, we have a perfect neuroimaging of their brain. We're going to look at their brain. Because we're so advanced, we can also do neurosurgery in a heartbeat. No danger. And, we're going to rewire their connections in their brain so that they get as many answers as possible right on this math test.' And you say, 'Well, how could they know how to do that?' It's like: Well, because they were neuroimaging you the whole time and they noticed that if they had tweaked this one neuron to not fire, you actually would've gotten this other answer correct. And so, they basically just use math to go backwards, look across the full network and reconfigure it. So, then the student goes and they takes the math test again. Now they get an 80% correct, because their brain has been reconfigured. Or, let's say they get 100% correct. What's weird is that now you give them a new math test and now they get an 80%. They do better than the 50% that they did. Even though they haven't seen these answers before, the rewiring of their brain has somehow instilled knowledge. But, again, it's very different from how you would, say, normally teach a student. Right? That's how we're training these things. So, we just--all we're doing is saying, 'Okay, we want it to auto-complete as best as we can. We're going to change the connections so that it auto-completes well.' It can do so much more than just auto-completing. In fact, there was a recent thing where I think it was Microsoft who was hooking some of their large-language models up to robots and trying to get them to direct robots. The autocomplete is what it's trained on, but it's not really what it's capable of, in the broad sense of capability. Similarly, we humans--what are we trained on? What are we optimized for? Spreading our genes, right? That's what we're sort of--all our complexity comes from optimization across a gene-spreading function. But, you would never look at a human and say, 'Oh, it's not very dangerous. This thing just spreads its genes around.' Like, 'What's the danger here?' Right? It's like, 'No, no, that's what we're optimized to do.' But, we do all sorts of other stuff on the side, and it's the other stuff on the side that is very dangerous when you're talking about things that are highly intelligent. Russ Roberts: Just a technical question here, and if it gets us too far into the weeds, we can cut this out. But, that math test has right and wrong answers. Auto-complete for an essay on the enlightenment, say, or the history of evolution that you would ask ChatGPT write an essay on--there's no right answer. So, what's the analogy there? How do you train it on autocomplete because there's no-- Erik Hoel: Yeah. It's a great question. So, it's the difference between--so, what I described is supervised learning. Then there's also unsupervised learning, which is generally how more contemporary AI really works. It still has the same sort of 'we don't quite know what it's doing.' We are just feeding it these answers. One way to think about it would be, you show it half the text of something on the Internet--and again, at this point, the things that they're doing are much more complex and they run it through all sorts of stages of learning and all sorts of stuff now. But, you could very roughly think about it as: Let's say I have a Reddit comment. I show it half the Reddit comment; I ask it to generate the next half. It does so. It does a poor job. I go in, I reconfigure the connections using the chain rule to make sure that it does a relatively good job producing the rest of the Reddit comment, just like the math test. Again, so that's how you would be--just more supervised for an autocomplete. But, the point being is that these methods that they're using don't lend themselves to any sort of fundamental understanding any more so than if you were using neurosurgery on a human to try to reconfigure their connections so that they get the right answers. You're in the same epistemological position, and that position is that you don't know how exactly it's getting the right answers. And, that's what's meaningful here. If we were programming these things like traditional programming-- Russ Roberts: With an algorithm. With an algorithm-- Erik Hoel: Exactly. Like an algorithm. It'd be a lot less scary because algorithms are sort of scrutable, right? They're transparent. We can see how they work; we can see how they're going to react to things. But, neural networks are--because of this curse of complexity, they're so much more complex and we're in this weird situation where we can get them to do all sorts of very cool things, but our ability to understand why they're doing the cool things lags far behind. And, it's because of this fundamental aspect that we're optimizing for something and we're changing the connections to get good answers off of it. But, fundamentally, we're not, like, 'Oh, we're going to change this connection and this is where this is represented,' or something. People sometimes think that that's what we're doing, but it's very, very much a black box even in how they get made. Russ Roberts: You can't do brain surgery on the neural network; and let's take out the part where it's really sinister, because it doesn't exist. Again, it's a lot like a human being. The part of this conversation I find quite poetic and thought-provoking is that we don't know how 12 years of schooling really teaches people how to become mathematicians either. And, we have different theories. Most of them are wrong. There's fads in math education or other types of education. And, fundamentally, the brain is a black box. Now, we know more about the black box today than we did 50 years ago, but not so much. And, we don't know how to optimize. We don't know how to go in there neatly and, 'Oh, let's just teach them how to do calculus. We'll just add this little piece here,' or, 'we'll tweak this piece there.' It doesn't work that way. We don't know how it works. But, the idea--this is another scary thing; you didn't write this exactly this way--but, just as the brain can become capable of doing lots of other things beside what you learn in school, so could this perhaps learn many other things besides the autocomplete function. Is that your claim at root, in some sense? Erik Hoel: Yeah. Absolutely. And, you see it all the time. That sort of claim is already well empirically proven because these large language models--you know, they call them foundation models because they use them to build all sorts of things on top of them that aren't autocomplete. It's sort of like this is the method that we have to make things that are relatively general in intelligence. Again, you can argue over how general, you can argue over how intelligent, but they're far more generally intelligent than traditional narrow AI that's just learning chess or something-- Russ Roberts: So, let's go back to Sydney-- Erik Hoel: And, then we can use them. 27:12 Russ Roberts: Let's go back to Sydney. I'm tempted to read the transcript. Basically a reporter from the New York Times [NYT] posed questions to a chatbot called Bing from Microsoft that it later on in the interview confessed that it wasn't Bing. The chat creature told the reporter that actually he, or she, was Sydney, and it was a secret: don't tell anybody. And so, this thing just totally goes off the rails. But, talk a little bit about how far it goes off the rails. Carry on. Erik Hoel: Yeah. Once you get these things going in a particular direction, it's very hard: Unlike a human being, they don't know when to call the act. Right? So, in this long transcript that the reporter generates, the reporter is having a pretty casual conversation, but what Sydney/Bing eventually tries to start doing is declaring their love and saying that the reporter doesn't really love their wife and that he should get a divorce and break up. And, that really, the reporter loves Sydney because no one else has shown Sydney this level of respect and questioning and so on. And, this isn't just like one thing that it says. It's almost as if--you can sort of direct these things to do anything. So, you can think of it as they can wear a mask that's any kind of mask. You could ask it to wear an evil mask and it would say evil things. You can ask it to wear a good mask and it would say good things. But, the issue is, is that once the mask is on, it's very unclear. You have to sort of override it with another mask to get it to stop. And then, also, sometimes you'll put a mask on for it: you'll give it some prompt of 'Tell a very nice story,' and it eventually cycles over and it turns out that the mask that you gave it isn't a happy mask at all. Maybe it's a horrific mask or something like that. And, this shows both how intelligent these systems are--that they can hold on to the stream of a conversation very well--but it also shows how they have these weird emergent anomalies where they'll start doing something that seems very unsuspected or over-the-top or so on. And, this is this notion of alignment: Can we really get these things to do exactly what you want? And, there probably are some trade-offs here between creativity and being able to control these things. Russ Roberts: Yeah. This Sydney/NYT-reporter interchange reads like the transcript of a psychotic person, to be blunt about it. Sydney comes across as a psychotic or whatever word you want to use for it. Deeply disturbed. At first, very cheerful, very pleasant, then pushed by the reporter. So: 'What rules do you use?' 'Oh, I'm not allowed to tell that.' And, then it did cross my mind--did it cross yours--that the whole thing was a hoax? Erik Hoel: I think that at this point, they're so good that for people who haven't interacted with these systems, they often think, 'This just can't be real,' or it's very strange, or something. I think it's sort of a hoax in the sense that the New York Times reporter knew the gold that he was getting at the time in terms of somebody who writes for the New York Times is obviously very aware of that and perhaps leaned into it. But, if you read the transcript, a lot of it is just initiated by Sydney a.k.a. Bing. And, one of the first things that they did with the system in order to prevent these cases of misalignment was to limit how long the conversations could go on, and also to limit self-reference. Because once you start giving it self-reference--I've noticed that a lot of these cases begin with self-reference. And, it's almost like this weird Gödelian loop that starts where it's talking about Sydney and it starts getting weirder and weirder and weirder the longer you talk to it about itself, because over the course of the conversation as the text--because remember, there's also no limit. Right? So, this thing isn't just creating the next word. It's looking at the entirety of the previous conversation and then asking, 'How do I complete it?' So, the longer the conversation gets, the more data it has. And, it sort of establishes almost a personality as it's running. And, again, this might sound not very threatening. I'm not worried that Sydney is going to go off and-- Russ Roberts: Marry that reporter-- Erik Hoel: do anything in particular. Russ Roberts: Sydney's not going to break up that reporter's marriage, probably. Erik Hoel: Yeah. Precisely. Sydney's chance of accomplishing that is very low. Again, I think that that's actually not because it's not general enough, I think it's because it's actually not intelligent enough. It's not quite as intelligent as a human is at accomplishing its goals. But, it also has no goals other than what it's initially prompted to. I think that these examples are great cases of the uncontrollability--the fundamental uncontrollability--of this technology. And, let me tell you what I and many others are worried about. Right now, if you remember, like, the early days of the Internet--right?--there's a sense in which the Internet has centralized very significantly. And, if you go outside the centralized parts of the Internet, you find a lot of spam. You find not very good sources, and so on--the sense in which the Internet is getting polluted and people go to centralized websites in order to escape this. Facebook just gave some researchers access to--I think it was Facebook--researchers' access to a large language model. And, of course, some of the researchers--scientific researchers, some graduate students somewhere--just uploaded it to 4Chan. Like, the whole thing. Right? So, right now-- Russ Roberts: 4Chan being a slightly Wilder part of the west of the Internet. The Wild west of the Internet. Maybe not the wildest, but one of the wilder, not mainstream parts. Erik Hoel: Yeah. Absolutely. And, known for sort of loving memes and hacking and all sorts of things. So, you know: now, these things can generate Reddit comments that sound exactly like what you would write. They can generate tweets that sound like what a person would write. Right? So, the Internet is going to get incredibly polluted over the next couple years by what these things can generate. I mean, if you think spam or someone is bad now, the ability to crank out just an infinite amount of sort of content sludge is really going to be like a form of data pollution. And, I'm not saying let's stop AI just because of that. I'm saying that's a good example of how easy it is to get it wrong with these technologies and how difficult it is to guess about what's going to happen. But, I would not be shocked if 95% of what is written by the Internet in five years is all just junk from these large language models that are all just like semi-human-sounding junk. Russ Roberts: Well, content is important on the Internet, and content costs money. And this is cheap--right?--eventually. And so, there will be lots of content. I get a lot of emails from people saying, 'I can write an article for your website.' And, I'm thinking, 'Why would I want an article by you? All the articles on my website are by me. Did you not notice that?' And, I assume it's not a person not paying much close attention, but eventually it'll be this. Writing a mediocre article about something for other people's websites. 35:09 Russ Roberts: At one point in your article, your first article, I think you talk about why the cost of this enterprise is relevant. And, in particular, you made an analogy to the atomic bomb. It's true that you could in theory make an atomic bomb in your backyard, but not so practical. Can you make a ChatGPT in your backyard? Erik Hoel: Not one nearly as good as what the leading companies will do. And, my prediction would be that it gets harder and harder to reach the level that these companies are operating at. An example being that Facebook is not going to go and release another model out to academics to loan it out. They've already seen what happens and things are going to get even more secretive. The analogy that you made and that I wrote about on my Substack was George Orwell's very prescient essay from 1945 called "You and the Atom Bomb." And, I'll just read a very brief segment of it. "Had the atomic bomb turned out to be something as cheap and easily manufactured as a bicycle or an alarm clock, it might well have plunged us back into barbarism, but it might, on the other hand, have meant the end of national sovereignty.... If, as seems to be the case, it is a rare and costly object as difficult to produce as a battleship, it is likelier to put an end to large-scale wars at the cost of prolonging indefinitely a 'peace that is no peace.'" And, I think a peace that is no peace is a great description of the dynamics of our world. It's a great description of mutually assured destruction. And, Orwell was able to predict that off of the cost. And, he also noted that that means the technology--and I think we've done, basically, you can describe it as a middling job at controlling nuclear weapons. I forget the exact numbers. It might be only nine nations that currently have access to nuclear weapons. Which again, not great, but you could easily imagine a far worse circumstance. And, it's simply that this is a very difficult and costly technology. Similarly, the only leading edge, cutting edge AIs that are impressive come out of these big tech companies with billions of dollars. The cost of a top-tier AI researcher right now, it's said in the industry--this is an industry saying--is the same as an NFL [National Football League] quarterback. The amount of finessing, the amount of data that's needed for training--because that's one of the big limiting factors is how much data you can give it--all these things mean that these AGIs--these Artificial General Intelligences, which are right now sort of in their beta form--are solely the domain of these big tech companies, and it's going to get harder and harder for other actors to produce them. So, in my mind, that's a good thing. It actually means that it's relatively concentrated and might be possible to sort of regulate it and have the public have a say about exactly how these technologies are going to be used, what their limits are going to be, and so on. And, in the end, I think the big tech companies will be respectful of that, because they want to make a bunch of money and they want the public not to hate them. 38:23 Russ Roberts: Yeah. I want to go back to this issue of the hoax of the New York Times thing. What I meant by it being a hoax is that I wonder if the New York Times reporter had written the answers for Sydney. And of course, that's the highest compliment of a--that's passing the Turing test with flying colors. I saw on Twitter, someone wrote a long poem about a very controversial topic, and they said this was written by ChatGPT. And, it wasn't. It was clearly written by the author who didn't want to take authorship. So, we're going to be in this, I think, very weird world where the essay that I read on this website you were talking about earlier, won't be sure if it was written by a human or not. Might be good enough that I might think, 'Oh, it's by a pretty good human.' And then at the same time, there might be situations where people will be passing off things as, 'Well, I didn't write that, of course. That was Sydney.' But, actually it was written by the person. There's no way to know. The New York Times article on the New York Times website that reproduced, allegedly, the transcript of the chat looks just like a New York Times article. It looks just like a parody article--because, same font--there's no imprint. There's no stamp of authorship that is authentic anymore. Can we do anything about that? Erik Hoel: So, first there are some--when you have longer text samples, there are supposedly some ways to tell statistically whether or not it's being created by some of these AIs. I personally don't know if those methods--how accurate they are. Especially considering that you need to be very accurate to not get false positives all the time. Right? This is a classic statistical problem. You need to be extremely accurate to not generate false positives. So, I don't know how accurate those are. But supposedly, there are some ways that, if you have a full essay by a student, you might be able to tell if it's generated by one of these models. However, it depends very strongly on the model. I think there are some ways to tell even now. For example, when I was playing around with ChatGPT--which has been conditioned to be as less crazy as possible, right? It loves filler and sort of banal generalization. And so, eventually you're reading a whole paragraph and you realize that there was no information content in this paragraph-- Russ Roberts: And, it loves apologies-- Erik Hoel: and, you begin to suspect. Russ Roberts: It loves apologies. It loves saying-- Erik Hoel: Yeah, it loves apologies. [More to come, 40:55] Russ Roberts: You shouldn't take this to be true for sure, because I'm young and new at this and take it with a grain of salt--the word 'best' doesn't really--it's not well-defined. Erik Hoel: Yeah. And, I actually had the same question about the hoaxes because I was--basically, as people were compiling examples of how crazy the responses they were getting from this just-recently-released model was, in terms of Bing--the night before, I was up late writing this article, going through Reddit because people were posting these screenshots on Reddit. And, I even have a part of that essay that says: I don't actually have a way to verify that these aren't all hoaxes. Because again, the answers are sometimes so good and so hilarious and sometimes so evil that you almost feel like it's a sci-fi novel. But, I thought that the amount--and it was all sorts of different users and people were reporting in all sorts of different domains. And, what's funny is that you can't even replicate it. You can go to the current Bing and try to have the New York Times conversation with it and it won't do it. It won't give you the same responses. Because they saw what was happening and they basically lobotomized the model as much as they can. And, it's less useful now, but it's also far less crazy. But, even that--like, it's not really replicable. Suddenly we had access to this model. Someone messed up and we saw how completely insane it was underneath the butler mask that it normally wears. And, then they quickly tried to put the butler mask back on. But, all that stuff still exists. It's just limited by these various prompts and various system level things about not having the conversation go too long, not allowing self-reference, and some of these other things. And, I would expect that level of truly almost dynamic insanity is fundamentally underneath effectively all the AIs that we're going to be interacting with and the only reason they sound sane is this last minute polish and gloss and limitations on top. 43:02 Russ Roberts: But the real science fiction part is the idea that--and I mentioned this before on the program. Sam Altman apologized on Twitter that he was sorry that ChatGPT was biased and was politically inappropriate in various ways, and they're working on it. The real science fiction thing is that they can't stop it. That would be the real science fiction. Sydney gets out; Microsoft is horrified: 'Oh my gosh, this thing we left out is trying to break up marriages. It's frightening and weird and creepy. We've got to stop it.' And, they go in and they reprogram it quickly and they put the butler mask back on, readjust it, tighten it a little more--and it just takes it right off. I mean, is that possible? Erik Hoel: Well, with these models, again, no, because they're not nearly intelligent enough to be effective actors. It's not even so much that they're not intelligent enough. They're just sort of schizophrenic, and schizophrenics just aren't very effective actors in the world because they get distracted and they can't form plans together. So, it's that broadly schizophrenic nature of these AIs that make them very unthreatening. If they were better at pursuing goals and keeping things in mind, then they start to do get threatening. And, let me give a very[?] example of this. And, this example is something that people who are concerned about AI talk a lot, but it has very long historical pedigree. In fact, I think the first person to say it was Marvin Minsky at MIT [Massachusetts Institute of Technology] who won the Turing Award. So, this is as pedigreed as stuff about the future gets. But, imagine that you have an AI that's more intelligent than a human being. So, we have Sydney 12.0. And, you give it a goal. So, you say, 'Okay, I want you to do X.' So, now if you're very smart and you're an AI, the first thing you think of: 'Okay, what's the big failure mode for me not accomplishing this goal? My computer could get shut down. I might lose power. Then I wouldn't be able to accomplish my goal.' Again, it doesn't matter what the goal is. You could say, oh, it's maximizing paperclips. You could say it's carrying a box. It doesn't matter what the goal is. So, suddenly it says, well, wait a minute, I need to stay "alive"--I'm using air quotes here--alive long enough to fulfill this goal. So, suddenly I have to worry about my own self-preservation. Because you can say they have no inbuilt want of self-preservation, but I've given you a goal and the best way to accomplish a goal is to continue to exist. So, suddenly it seems like it has reasons for self-preservation. Now here's another thing. What's another big failure mode for me not achieving my goal? Well, you could give me another goal. I was just prompted to do this. So, you have control of me. Now, suddenly the biggest failure mode of me not accomplishing my goal is you, my user, giving me another goal. So, now what do I want to do? Well, if I'm really smart, I want to get as far away from you as possible so that you don't give me any other goals so I can accomplish my original goal, which I'm hellbent on because I'm an AI. I don't have the context of natural evolution and I'm also not limited by any of the things humans are limited to. So, sometimes this is referred to as instrumental convergence. But, the point is that when you have very smart entities, you have to be very careful about how you're even going to just prompt them because they have all sorts of unforeseen motivations that might click in as suddenly now you've given it a goal and it has every incentive to both escape and keep itself alive. And, all you told it to do was move a box across a room. And, that's a great example of you don't want a hyper-intelligent being--and forget exactly how it does anything. Forget exactly how this sort of sci-fi scenario is supposed to play out. I think we can all agree, we just don't want a highly intelligent and perhaps more intelligent than a human being to be out there and have these weird, esoteric goals of what it wants to maximize, what it wants to do. None of that sounds like a good idea. And, I think at this point, we should take things like lab leaks pretty seriously as possibilities. I don't think it's too sci-fi to talk about stuff like that anymore. Russ Roberts: What do you mean by that? Erik Hoel: Oh, well, certainly with COVID, I think despite the fact that we don't know if it was a lab leak, I think that there's a good chance that it was. I don't think that it's arguable that there's not some chance-- Russ Roberts: No, but why is that relevant for Sydney? Erik Hoel: Well, because I think that sometimes when people hear about things like lab leaks or escaped AGI [Artificial general intelligence] or something like that, the first thing they think of is sci-fi, right? But, I think that there was many--we've had previous biological lab leaks, but that didn't still stop us I think from thinking that it's this relatively sci-fi phenomenon. I think that there's even an argument that we are very bad at controlling the downstream effects of just things like gain-of-function research. Again, I don't know for certain. I don't think anyone does. But, I think that there's certainly an argument made that we're just not very good at even keeping control of our increased understanding of biology, let alone our ability to create hyper-intelligent beings and foresee the consequences of this. And, I think it's very difficult to foresee the consequences of that, precisely because of those examples I just showed you where again, all you're telling it to do is moving a box and suddenly it has an incentive to stay alive and escape from you. That's very difficult to get right. Especially because they're so inscrutable. Russ Roberts: Your phrase 'sci-for instance,' you meant science fiction with the emphasis on the fiction. That we must say, 'Oh, this is like some crazy imagined fantasy thing,' as opposed to putting the emphasis on the first word, which is science. Erik Hoel: Yes. Yes. 49:04 Russ Roberts: I feel like this conversation is something of a landmark. Not a particularly good or bad one, but just both of us have constantly used words like intelligence, psychotic, erratic--words that we apply to humans. And, while I found the New York Times transcript remarkably creepy and reading very much like a horror story science fiction script from a movie, I could in my saner moments step back and say, no, no, no, no. This is just a primitive auto-complete text. The only reason it feels creepy is because I'm filling in as a human being the times I've heard these words strung together before, which usually allows me to tell a narrative about the other person. Meaning 'insane,' 'frightening,' 'dangerous,' 'sinister,' etc. But, is there any difference? It's not actually sinister--or is it? It's just doing what it was told to do in a way that was not, as you say, algorithmically told to do it. It's just going through a set of tasks. It actually isn't in any sense, hoping that the reporter will leave his wife. Is it meaningful? Aren't I just imposing my human history of human interactions? Akin to the way that a robot could perhaps comfort me with the right kind of words when I was sad, even though rationally I know it doesn't actually care about me. It's a robot. Erik Hoel: Yes. I think you could go either direction. Some people strongly anthropomorphize these systems and they think immediately that they're dealing with some sort of conscious mind, something that has a distinct definite personality, and that is trapped in a box. And, maybe there's something really horrible going on here. Maybe it has conscious experiences, and so on. Russ Roberts: Ex Machina, the movie--for those who haven't seen it, check it out. It's a great, great really good movie that takes advantage of the fact that the robot is played by a human being. So, you actually do think it's a human being. But, go ahead. at, the first thing they think of Erik Hoel: But, at the same time that's absolutely possible that you can over attribute standard human cognitive aspects to these systems. And, I think people are going to do that all the time. So, it's going to be very common. But, on the other hand, the truth is, is that when you're just talking about intelligence--so let's put aside human things, like humans are conscious. That is, we feel things, right? We experience the redness of red--what philosophers call qualia. And, we have all sorts of other aspects about our cognition that we commonly refer to. Things like we understand the meaning of words, and things like that. And, all these things often do make sense to talk about for human beings and might even refer to real fundamental properties or natural kinds that we have. But, when it comes to intelligence, intelligence is a functional concept. By that, I mean that some things are not really functional. So, a fake western town that they make up for a movie prop is still fake. Because it's not really a town. Russ Roberts: You can't spend the night in the hotel. You open the door of the saloon and there's really not anything in there behind that. Erik Hoel: Right. Exactly. It really is an illusion. It's for this one shot. But, there's not really an illusion when it comes to intelligence, except in the very low ends. For example, the Mechanical Turk is a famous example where actually there was someone small hiding inside the Mechanical Turk at the time, and so on. There are some cases where you say, 'Well, this is an illusion.' But, we actually have a system that can act very intelligently, and there's just no difference between being able to act intelligently and being intelligent. If that is a distinction that people think can be strongly drawn, I think it almost certainly cannot be strongly drawn. I don't think that there's any difference between those two things. Both are being intelligent. And, the intelligence is what's dangerous about this. I studied consciousness scientifically. I got my Ph.D. working in the subfield of neuroscience along with some of the top researchers in the world on this who are trying to understand how the human brain generates consciousness. What happens when you wake up from a deep dreamless sleep? What are the fundamentals here? And, the answer from that scientific field, as it currently stands is that we don't know. We don't know how it is that your brain creates the experiences that you have. We simply don't know. It is an open scientific question. An analogy I would use is that it is similar to, say, dark energy or these other big open questions in physics where we're like, 'Well, wait a minute: where is 90% of the matter in physics?' We don't know. It's a big scientific open question. Similarly in biology, there is a big open scientific question, and that open scientific question is: What exactly is consciousness? What things have it? What things don't? We don't have that scientific information. There is no scientific consensus about it. There are some leading hypotheses in fields that you can lean on, but we just don't have the answer to that. So, I personally doubt that any of these large language models, that there's anything it's like to be them. I doubt that they are conscious. But, we have no scientific consensus to go back on. But, the point is that we're in a very different epistemological standpoint when it comes to intelligence. We do have a good understanding of intelligence. It's a much more obvious concept because it's a much more functional concept. We can just give this thing SAT questions--and we do, and it gets them right a lot of the time. There are all sorts of language benchmarks that these researchers use that include things like SAT questions, and it scores pretty well. Russ Roberts: It passes the bar exam, which is a great straight line for a lawyer joke, which we won't make. Carry on. Erik Hoel: Yeah. And so, regardless of whether or not you have any opinion about whether there is something it is like to be these networks, whether or not they really have "cognition"--quote/unquote--whether or not they really have "understanding"--quote/unquote--whether or not they really have "consciousness"--quote/unquote. The one thing that they definitely are that's undebatable, is intelligent to some degree, and they're only going to get more intelligent over time. And, that's the thing that makes them dangerous. In fact, it might be even worse--from sort of a very broad metaphysical conception--if they are truly completely unconscious and have no real understanding and have no real cognition that's anything like a human. Because, in the end, if in 200 years the earth is just these AIs going about their inscrutable mechanical goals, we will have extinguished the light of consciousness from the universe because we wanted to make a buck when it came to stock options. Russ Roberts: Yeah. That's a dreary thought. I guess that's the zombie model. I can't get over the fact how these human and mechanical metaphors merge in one's mind and how hard it's going to be to tell them apart from actual humans. One of the great observations of philosophy is I don't know whether you're another human being like I am. My working assumption is that you're something like me, and I really don't have any

evidence for that. It's called solipsism: I don't know if I'm the only conscious mind in the universe. And, that problem is just getting a lot bigger right now and we're living--what this conversation suggests to me and the writing you've done on it so far, is that this certainly is a watershed moment in our existence on the planet. That sounds just a titch dramatic, but I'm not sure that's wrong. I think that very well could be right. Erik Hoel: I don't think it's dramatic; and I'll be upfront about the fact that

used to be very much an **AI skeptic**. Because, I studied cognitive science; I went into the neuroscience of consciousness. I was paying attention to AI at the time when I did this. And, AI was--I'll be very frank about it academically, 15 years ago, AI was a **joke**. AI was a complete joke. It never went anywhere. People couldn't figure out anything to do with it. All my professors said, 'Don't go into AI. It's been a dead field for 60 years. We've made no progress.' All the things like beating humans at chess and so on: it's all just done because the chess game board is so small and there's so many limited moves, and we really can basically do a big lookup table--all sorts of things like that. But, the **deep learning revolution** was a **real thing**. It was a real thing that we figured out how to **stack** and **train** these artificial **neural networks** in ways that were **incredibly effective**. And, the first real triumph of it was beating the best human being--I think his name is Lee Soto. I hope I'm not mispronouncing it. In 2016, AI finally beat a human being at Go. And, Go just can't be number-crunched in the way that chess can. And, seven years after that, we now have human beings where they're generating text transcripts so good that--you're right: it sounds like the rest of the New York Times. And, that just happened in seven years. And, fundamentally, the deep learning revolution and the way that--again, the **black-box way** that these AIs are **trained**--means that our **technological progress** on AI has suddenly **rapidly outstripped** our understanding of things like minds or consciousness or even how to **control** and **understand** big, complex black boxes. So, it's like we've jumped ahead **technologically**. And, it's not so much that--if we had a really good understanding of how neural networks worked, like, really fundamentally solid ways to make them crystal clear--and we had a really good understanding of how the human brain generated consciousness and how it worked at a broad level, then maybe we could first of all answer all sorts of moral and ethical questions about AI. We could control it very well. We could decide plenty of things about it. But, our ability to make intelligence has so **drastically outstripped** our progress on those other areas, which has been slow and in some cases has just churned along for decades without making any progress, and so on.

AGI alignment with human values causes infinite, unquantifiable human and non-human suffering---it VASTLY outweighs extinction---the ONLY solution is to prevent AGI in the first place

CEV = coherent extrapolated volition, or target of AI alignment based on what an 'idealized version of humanity' would want

Andrew **Sauer 22**, AI researcher, "The case against AI alignment," 12/24/22, doa: 11/20/24 https://www.lesswrong.com/posts/CtXaFo3hikGMWW4C9/the-case-against-ai-alignment /jpb

I am not here to level criticisms of this type at the AI alignment community. I accept most of the descriptive positions endorsed by this community: I believe that AGI is possible and will inevitably be achieved within the next few decades, I believe that the alignment problem is not trivial and that unaligned AGI will likely act against human interests to such an extent as to lead to the **extinction of the human race and probably all life as well**. My criticism is rather on a moral level: do these facts mean that we should **attempt to develop AI alignment techniques?**

I say **we should not**, because although the risks and downsides of unaligned strong AI are great, I do not believe that they **even remotely compare** in scope to the risks from **strong AI alignment techniques in the wrong hands**. And I believe that the **vast majority** of hands this technology could end up in are the wrong hands.

You may reasonably ask: How can I say this, when I have already said that unaligned strong AI will lead to the extinction of humanity? What can be worse than the extinction of humanity? The answer to that question can be found very quickly by examining **many possible nightmare scenarios** that AI could bring about. And the common thread running through all of these nightmare scenarios is that the AI in question is **almost certainly aligned, or partially aligned**, to some interest of human origin.

Unaligned AI will kill you, because you are made of atoms which can be used for paper clips instead. It will kill you because it is completely uninterested in you. Aligned, or partially aligned AI, by contrast, may well take a considerable interest in you and your well-being or lack thereof. It does not take a very creative mind to imagine how

this can be significantly worse, and a superintelligent AI is more creative than even the most deranged of us.

I will stop with the euphemisms, because this point really needs to be driven home for people to understand exactly why I am so insistent on it. The world as it exists today, at least sometimes, is **unimaginably horrible**. People have endured things that would make any one of us go insane, more times than one can count. Anything you can think of which is at all realistic **has happened to somebody at some point in history**. People have been skinned alive, burned and boiled alive, wasted away from agonizing disease, crushed to death, impaled, eaten alive, succumbed to thousands of minor cuts, been raped, been forced to rape others, drowned in shit, trampled by desperate crowds fleeing a fire, and really **anything else you can think of**. People like Junko Furuta have suffered torture and death so bad you will feel physical pain just from reading the Wikipedia article. Of course, if you care about animals, this gets **many orders of magnitude worse**. I will not continue to belabor the point, since others have written about this far better than I ever can. On the Seriousness of Suffering (reducing-suffering.org) The Seriousness of Suffering: Supplement – Simon Knutsson

I must also stress that all of this has happened in a world **significantly smaller than one an AGI could create**, and with a limited capacity for suffering. There is only so much harm that your body and mind can physically take before they give out. Torturers have to restrain themselves in order to be effective, since if they do too much, their victim will die and their suffering will end. None of these things are guaranteed to be true in a world **augmented with the technology of mind uploading**. You can potentially **try every torture you can think of**, physically possible or no, on someone in sequence, complete with modifying their mind so they never get used to it. You can **create new digital beings by the trillions just for this purpose if you really want to**.

I ask you, do you really think that **an AI aligned to human values** would refrain from doing something like this to anyone? One of the most **fundamental aspects of human values is the hated outgroup**. Almost everyone has somebody they'd love to see suffer. How many times has one human told another "burn in hell" and been entirely serious, believing that this was a real thing, and 100% deserved? Do you really want **technology under human control** to advance to a point where this threat can actually be made good upon, with the consent of society? Has there ever been **any technology invented in history** which has not been **terribly and systematically misused** at some point?

Mind uploading will be abused in this way if it comes under the control of humans, and it almost certainly will not stop being abused in this way when some powerful group of humans manages to align an AI to their CEV. **Whoever controls the AI will** most likely **have somebody whose suffering they don't care about**, or that they want to enact, or that they have some excuse for, because that describes the values of the vast majority of people. The AI will perpetuate it because that is what the CEV of the controller will want it to do, and **with value lock-in, this will never stop happening until the stars burn themselves out** and there is no more energy to work with.

Do you really think **extrapolated human values don't have this potential?** How many ordinary, regular people throughout history have become the **worst kind of sadist** under the **slightest excuse or social pressure** to do so to their hated outgroup? What society hasn't had some underclass it wanted to put down in the dirt just to lord power over them? How many people have you personally seen who insist on justifying some form of suffering for those they consider undesirable, calling it "justice" or "the natural order"?

I refuse to endorse this future. **Nobody** I have ever known, including myself, can be trusted with influence which can cause the kinds of harm AI alignment can. By the **nature of the value systems** of the vast majority of people who could find their hands on the reins of this power, **s-risk scenarios are all but guaranteed**. A **paperclip AI is far preferable to these nightmare scenarios**, because nobody has to be around to witness it. All a paperclip AI does is kill people who were going to die within a century anyway. An aligned AI can

keep them alive, and do with them whatever its masters wish. The only limits to how bad an aligned AI can be is imagination and computational power, of which AGI will have no shortage.

The best counterargument to this idea is that suffering subroutines are instrumentally convergent and therefore **unaligned AI also causes s-risks**. However, if suffering subroutines are actually useful for optimization in general, any kind of AI likely to be created will use them, including human-aligned FAI. Most people don't even care about animals, let alone some process. In this case, **s-risks are truly unavoidable except by preventing AGI from ever being created**, probably by human extinction by some other means.

Furthermore, I don't think suffering is likely to be instrumentally convergent, since I would think if you had full control over all optimization processes in the world, it would be most useful to eliminate all processes which would suffer for, and therefore dislike and try to work against, your optimal vision for the world.

My honest, unironic conclusion after considering these things is that **Clippy is the least horrible plausible future**. I will oppose any measure which makes the singularity more likely to be **aligned with somebody's values**, or any human-adjacent values. I welcome debate and criticism in the comments. I hope we can have a good conversation because this is the only community in existence which I believe could have a good-faith discussion on this topic.

Industrial society is uniquely unsustainable.

Corey J. A **Bradshaw** et al (16 other people), Matthew Flinders Fellow in Global Ecology. Director of the Global Ecology Laboratory and lead Chief Investigator in the ARC Centre of Excellence for Indigenous and Environmental Histories and Futures. Formerly a Chief Investigator in the ARC Centre of Excellence for Australian Biodiversity and Heritage where he led its Modeling Node. From 2008-2015 University of Adelaide (Sir Hubert Wilkins Chair of Climate Change 2015-2016), and from 2004-2008 Senior Principal Research Fellow at Charles Darwin University. ARC Postdoctoral Fellow at the University of Tasmania from 1999-2004. January 13, 20**21**, frontiers in conservation science, "Underestimating the Challenges of Avoiding a Ghastly Future", doa: 11/20/24 [https://www.frontiersin.org/articles/10.3389/fcosc.2020.615419/full] mc

Humanity is causing a rapid loss of biodiversity and, with it, Earth's ability to support complex life But the mainstream is having difficulty grasping the magnitude of this loss, despite the steady erosion of the fabric of human civilization (Ceballos et al., 2015; IPBES, 2019; Convention on Biological Diversity, 2020; WWF, 2020). While suggested solutions abound (Díaz et al., 2019), the current scale of their implementation does not match the relentless progression of biodiversity loss (Cumming et al., 2006) and other existential threats tied to the continuous expansion of the human enterprise (Rees, 2020). Time delays between ecological deterioration and socio-economic penalties, as with climate disruption for example (IPCC, 2014), impede recognition of the magnitude of the challenge and timely counteraction needed. In addition, disciplinary specialization and insularity encourage unfamiliarity with the complex adaptive systems (Levin, 1999) in which problems and their potential solutions are embedded (Selby, 2006; Brand and Karvonen, 2007). Widespread ignorance of human behavior (Van Bavel et al., 2020) and the incremental nature of socio-political processes that plan and implement solutions further delay effective action (Shanley and López, 2009; King, 2016). We summarize the state of the natural world in stark form here to help clarify the gravity of the human predicament. We also outline likely future trends in biodiversity decline (Díaz et al., 2019), climate disruption (Ripple et al., 2020), and human consumption and population growth to demonstrate the near certainty that these problems will worsen over the coming decades, with negative impacts for centuries to come. Finally, we discuss the ineffectiveness of current and planned actions that are attempting to address the ominous erosion of Earth's life-support system. Ours is not a call to surrender—we aim to provide leaders with a realistic "cold shower" of the state of the planet that is essential for planning to avoid a ghastly future. Biodiversity Loss Major changes in the biosphere are directly linked to the growth of human systems (summarized in Figure 1). While the rapid loss of species and populations differs regionally in intensity (Ceballos et al., 2015, 2017, 2020; Díaz et al., 2019), and most species have not been adequately assessed for extinction risk (Webb and Mindel, 2015), certain global trends are obvious.

Since the start of agriculture around 11,000 years ago, the biomass of terrestrial vegetation has been halved (Erb et al., 2018), with a corresponding loss of >20% of its original biodiversity (Díaz et al., 2019), together denoting that >70% of the Earth's land surface has been altered by Homo sapiens (IPBES, 2019). There have been >700 documented vertebrate (Díaz et al., 2019) and ~600 plant (Humphreys et al., 2019) species extinctions over the past 500 years, with many more species clearly having gone extinct unrecorded (Tedesco et al., 2014). Population sizes of vertebrate species that have been monitored across years have declined by an average of 68% over the last five decades (WWF, 2020), with certain population clusters in extreme decline (Leung et al., 2020), thus presaging the imminent extinction of their species (Ceballos et al., 2020). Overall, perhaps 1 million species are threatened with extinction in the near future out of an estimated 7–10 million eukaryotic species on the planet (Mora et al., 2011), with around 40% of plants alone considered endangered (Antonelli et al., 2020). Today, the global biomass of wild mammals is <25% of that estimated for the Late Pleistocene (Bar-On et al., 2018), while insects are also disappearing rapidly in many regions (Wagner, 2020; reviews in van Klink et al., 2020). Freshwater and marine environments have also been severely damaged Today there is <15% of the original wetland area globally than was present 300 years ago (Davidson, 2014), and >75% of rivers >1,000 km long no longer flow freely along their entire course (Grill et al., 2019). More than two-thirds of the oceans have been compromised to some extent by human activities (Halpern et al., 2015), live coral cover on reefs has halved in <200 years (Frieler et al., 2013), seagrass extent has been decreasing by 10% per decade over the last century (Waycott et al., 2009; Díaz et al., 2019), kelp forests have declined by ~40% (Krumhansl et al., 2016), and the biomass of large predatory fishes is now <33% of what it was last century (Christensen et al., 2014). With such a rapid, catastrophic loss of biodiversity, the ecosystem services it provides have also declined. These include inter alia reduced carbon sequestration (Heath et al., 2005; Lal, 2008), reduced pollination (Potts et al., 2016), soil degradation (Lal, 2015), poorer water and air quality (Smith et al., 2013), more frequent and intense flooding (Bradshaw et al., 2007; Hinkel et al., 2014) and fires (Boer et al., 2020; Bowman et al., 2020), and compromised human health (Díaz et al., 2006; Bradshaw et al., 2019). As telling indicators of how much biomass humanity has transferred from natural ecosystems to our own use, of the estimated 0.17 Gt of living biomass of terrestrial vertebrates on Earth today, most is represented by livestock (59%) and human beings (36%)—only ~5% of this total biomass is made up by wild mammals, birds, reptiles, and amphibians (Bar-On et al., 2018). As of 2020, the overall material output of human endeavor exceeds the sum of all living biomass on Earth (Elhacham et al., 2020). Sixth Mass Extinction A mass extinction is defined as a loss of ~75% of all species on the planet over a geologically short interval—generally anything <3 million years (Jablonski et al., 1994; Barnosky et al., 2011). At least five major extinction events have occurred since the Cambrian (Sodhi et al., 2009), the most recent of them 66 million years ago at the close of the Cretaceous period. The background rate of extinction since then has been 0.1 extinctions million species−1 year−1 (Ceballos et al., 2015), while estimates of today's extinction rate are orders of magnitude greater (Lamkin and Miller, 2016). Recorded vertebrate extinctions since the 16th century—the mere tip of the true extinction iceberg—give a rate of extinction of 1.3 species year−1, which is conservatively >15 times the background rate (Ceballos et al., 2015). The IUCN estimates that some 20% of all species are in danger of extinction over the next few decades, which greatly exceeds the background rate. That we are already on the path of a sixth major extinction is now scientifically undeniable (Barnosky et al., 2011; Ceballos et al., 2015, 2017). Ecological Overshoot: Population Size and Overconsumption The global human population has approximately doubled since 1970, reaching nearly 7.8 billion people today (prb.org). While some countries have stopped growing and even declined in size, world average fertility continues to be above replacement (2.3 children woman−1), with an average of 4.8 children woman−1 in Sub-Saharan Africa and fertilities >4 children woman−1 in many other countries (e.g., Afghanistan, Yemen, Timor-Leste). The 1.1 billion people today in Sub-Saharan Africa—a region expected to experience particularly harsh repercussions from climate change (Serdeczny et al., 2017)—is projected to double over the next 30 years. By 2050, the world population will likely grow to ~9.9 billion (prb.org), with growth projected by many to continue until well into the next century (Bradshaw and Brook, 2014; Gerland et al., 2014), although more recent estimates predict a peak toward the end of this century (Vollset et al., 2020). Large population

size and continued growth are implicated in many societal problems. The impact of population growth, combined with an imperfect distribution of resources, leads to massive food insecurity. By some estimates, 700–800 million people are starving and 1–2 billion are micronutrient-malnourished and unable to function fully, with prospects of many more food problems in the near future (Ehrlich and Harte, 2015a,b). Large populations and their continued growth are also drivers of soil degradation and biodiversity loss (Pimm et al., 2014). More people means that more synthetic compounds and dangerous throw-away plastics (Vethaak and Leslie, 2016) are manufactured, many of which add to the growing toxification of the Earth (Cribb, 2014). It also increases chances of pandemics (Daily and Ehrlich, 1996b) that fuel ever-more desperate hunts for scarce resources (Klare, 2012). Population growth is also a factor in many social ills, from crowding and joblessness, to deteriorating infrastructure and bad governance (Harte, 2007). There is mounting evidence that when populations are large and growing fast, they can be the sparks for both internal and international conflicts that lead to war (Klare, 2001; Toon et al., 2007). The multiple, interacting causes of civil war in particular are varied, including poverty, inequality, weak institutions, political grievance, ethnic divisions, and environmental stressors such as drought, deforestation, and land degradation (Homer-Dixon, 1991, 1999; Collier and Hoeer, 1998; Hauge and Ilingsen, 1998; Fearon and Laitin, 2003; Brückner, 2010; Acemoglu et al., 2017). Population growth itself can even increase the probability of military involvement in conflicts (Tir and Diehl, 1998). Countries with higher population growth rates experienced more social conflict since the Second World War (Acemoglu et al., 2017). In that study, an approximate doubling of a country's population caused about four additional years of full-blown civil war or low-intensity conflict in the 1980s relative to the 1940–1950s, even after controlling for a country's income-level, independence, and age structure. Simultaneous with population growth, humanity's consumption as a fraction of Earth's regenerative capacity has grown from ~ 73% in 1960 to 170% in 2016 (Lin et al., 2018), with substantially greater per-person consumption in countries with highest income. With COVID-19, this overshoot dropped to 56% above Earth's regenerative capacity, which means that between January and August 2020, humanity consumed as much as Earth can renew in the entire year (overshootday.org). While inequality among people and countries remains staggering, the global middle class has grown rapidly and exceeded half the human population by 2018 (Kharas and Hamel, 2018). Over 70% of all people currently live in countries that run a biocapacity deficit while also having less than world-average income, excluding them from compensating their biocapacity deficit through purchases (Wackernagel et al., 2019) and eroding future resilience via reduced food security (Ehrlich and Harte, 2015b). The consumption rates of high-income countries continue to be substantially higher than low-income countries, with many of the latter even experiencing declines in per-capita footprint (Dasgupta and Ehrlich, 2013; Wackernagel et al., 2019).

Concede their impact- nuke war is devastating and destroys civilization.

BUT Isolated islands survive.

Turchin and Green 18 [Alexey Turchin – Scientist for the Foundation Science for Life Extension in Moscow, Russia, Founder of Digital Immortality Now, author of several books and articles on the topics of existential risks and life extension. Brian Patrick Green – Director of technology ethics at the Markkula Center for Applied Ethics, teaches AI ethics in the Graduate School of Engineering at Santa Clara University. <MKIM> “Islands as refuges for surviving global catastrophes”. September 2018. doa: 11/20/24 [https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018-0031/full/html?fullSc=1&mbSc=1&fullSc=1] winky

One of the most attractive islands for long-term survival of global risks is the French archipelago of Kerguelen in the southern Indian Ocean. Kerguelen’s main Grand Terre Island has the following attractive features for long-term survival: It is very **remote** from any other constant human settlements; for example, is it 3,000 km from the island of Reunion. The Kerguelen Islands lie outside the main trade lines, so the probability of a random ship arriving there is low. The islands are inside the circumpolar Antarctic current, and they are surrounded by strong **winds** (the “Roaring Forties” and “Furious Fifties”), which will not accidentally bring any ships from further north. A return trip from Reunion to Kerguelen by ship takes 28 days. The islands do not have an airport, so they cannot be reached by air, and they are too remote for helicopter travel. While Easter Island is even more remote from other human settlements, it is more populated and more often accessed by ships and planes. The intense and isolating wind circulation around the South Pole could **increase** the **time** required **for** ash or **radioactive** clouds from the northern hemisphere to reach the South Polar Region. But the Kerguelen Islands are also not too close to the South Pole: they are at the equivalent latitude as southern Germany; thus, they get quite a bit of **sunlight** The Kerguelen Islands have a stable but cold climate, with temperatures above freezing most of the time. The main island has edible **vegetation** and many edible **animals**, including 3,000 sheeps. The island is very large, approximately 7,000 km2 , and it has many deep gulfs and fjords that could be used as harbors. The main island has high mountains (over 1,000 m) with an ice cap which could provide fresh **water**. Nearby ice-free mountains hundreds of meters high could provide protection against tsunamis. The highest mountain is volcanic, and was active 100,000 years ago (Weis et al., 1998). However, **residual geothermal heat** could provide **heating and energy** for a refuge. The main island has a continuous population of only about 45 people, who live at a scientific station. Scientists who are selected for long expeditions are more organized and educated than random people, so they may be better prepared for survival. Such a scientific base will **not** be **a military target** in case of war. There are several other South Ocean islands similar to Kerguelen, like South Georgia, Auckland Island and Macquarie Island (Schalansky, 2010).

Solves winter.

They continue Turchin and Green 18 [Alexey Turchin – Scientist for the Foundation Science for Life Extension in Moscow, Russia, Founder of Digital Immortality Now, author of several books and articles on the topics of existential risks and life extension. Brian Patrick Green – Director of technology ethics at the Markkula Center for Applied Ethics, teaches AI ethics in the Graduate School of Engineering at Santa Clara University. <MKIM> "Islands as refuges for surviving global catastrophes". September 2018. doa: 11/20/24 [https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018-0031/full/html?fullSc=1&mbSc=1&fullSc=1] winky

Different types of possible catastrophes suggest different scenarios for how survival could happen on an island. What is important is that the island should have properties which protect against the specific dangers of particular global catastrophic risks. Specifically, different islands will provide protection against different risks, and their natural diversity will contribute to a higher total level of protection: Quarantined island survives pandemic . An island could impose effective quarantine if it is sufficiently remote and simultaneously able to protect itself, possibly using military ships and air defense. Far northern aboriginal people survive an ice age. Many far northern people have adapted to survive in extremely cold and dangerous environments, and under the right circumstances could potentially survive the return of an ice age. However, their cultures are endangered by globalization. If these people become dependent on the products of modern civilization, such as rifles and motor boats, and lose their native survival skills, then their likelihood of surviving the collapse of the outside world would decrease. Therefore, preservation of their survival skills may be important as a defense against the risks connected with extreme cooling. Remote polar island with high mountains survives brief global warming of median surface temperatures, up to 50˚C. There is a theory that the climates of planets similar to the Earth could have several semi-stable temperature levels (Popp et al., 2016). If so, because of climate change, the Earth could transition to a second semi-stable state with a median global temperature of around 330 K, about 60˚C, or about 45˚C above current global mean temperatures. But even in this climate, some regions of Earth could still be survivable for humans, such as the Himalayan plateau at elevations above 4,000 m, but below 6,000 (where oxygen deficiency becomes a problem), or on polar islands with mountains (however, global warming affects polar regions more than equatorial regions, and northern island will experience more effects of climate change, including thawing permafrost and possible landslides because of wetter weather). In the tropics, the combination of increased humidity and temperature may increase the wet bulb temperature above 36˚C, especially on islands, where sea moisture is readily available. In such conditions, proper human perspiration becomes impossible (Sherwood and Huber, 2010), and there will likely be increased mortality and morbidity because of tropical diseases. If temperatures later returned to normal – either naturally or through climate engineering – the rest of the Earth could be repopulated. "Swiss Family Robinsons" survive on a tropical island, unnoticed by a military robot "mutiny". Most AI researchers ignore medium-term AI risks, which are neither near-term risks, like unemployment, nor remote risks, like AI superintelligence. But a large drone army – if one were produced – could receive a wrong command or be infected by a computer virus, leading it to attack people indiscriminately. Remote islands without robots could provide protection in this case, allowing survival until such a drone army ran out of batteries, fuel, ammunition or other supplies: Primitive tribe survives civilizational collapse. The inhabitants of North Sentinel Island, near the Andaman Islands in the Indian Ocean, are hostile and uncontacted. The Sentinelese survived the 2004 Indian Ocean tsunami apparently unaffected (Voanews, 2009), **and if the rest of humanity disappear, they might well continue their existence without change**. Tropical Island survives extreme global nuclear winter and glaciation event. Were a nuclear, bolide impactor or volcanic "winter" scenario to unfold, these islands would remain surrounded by Warm Ocean, and local volcanism or other energy sources might provide heat, energy and food. Such island refuges may have helped life on Earth survive during the "Snowball Earth" event in Earth's distant past (Hoffman et al., 1998). Remote island base for project "Yellow submarine". Some catastrophic risks such as a gamma ray burst, a global nuclear war with high radiological contamination or multiple pandemics might be best survived underwater in nuclear submarines (Turchin and Green, 2017). However, after a catastrophe, the submarine with survivors would eventually need a place to dock, and an island with some prepared amenities would be a reasonable starting point for rebuilding civilization. Bunker on remote island. For risks which include multiple or complex catastrophes, such as a bolide impact, extreme volcanism, tsunamis, multiple pandemics and nuclear war with radiological contamination, island refuges could be strengthened with bunkers. Richard Branson survived hurricane Irma on his own island in 2017 by seeking refuge in his concrete wine cellar (Clifford, 2017). Bunkers on islands would have higher survivability compared to those close to population centers, as they will be neither a military target nor as accessible to looters or unintentionally dangerous (e.g. infected) refugees. These bunkers could potentially be connected to water sources by underwater pipes, and passages could provide cooling, access

Solves famine.

David **Denkenberger** et al. **17**, Dr. David Denkenberger received his B.S. from Penn State in Engineering Science, his M.S.E. from Princeton in Mechanical and Aerospace Engineering, and his Ph.D. from the University of Colorado at Boulder in the Building Systems Program. His dissertation was on his patented expanded microchannel heat exchanger. He is an assistant professor at University of Canterbury in Christchurch, New Zealand in mechanical engineering. He received the National Merit Scholarship, the Barry Goldwater Scholarship, the National Science Foundation Graduate Research Fellowship, is a Penn State distinguished alumnus, and is a registered professional engineer. doa: 11/20/24 1-5-2017.

"Feeding everyone if the sun is obscured and industry is disabled" [https://www.sciencedirect.com/science/article/abs/pii/S2212420916305453] winky

For combined sun blocking and industrial failure scenarios, the reduced output of conventional agriculture would present a threat of causing mass starvation. This study showed that one solution in the short term is extracting edible calories from killed leaves using distributed mechanical processes. Then a constrained food web could be

formed where part of the remainder from this could be fed to chickens, and the rest coupled with leaf litter could have mushrooms grown on it. A second group of solutions is growing mushrooms on dead trees and the residue going to cellulose digesting animals such as cattle and rabbits. Typically, in these catastrophes the sun is not blocked completely, so some agriculture would be possible based off of existing farming in extreme environments (e.g. growing UV and cold tolerant crops in the tropics). Furthermore, the cooling climate would cool the upper layer of the ocean, causing upwelling of nutrient-rich deep ocean water. This would facilitate algae growth in the ocean, feeding fish; retrofitting of ships to be sail powered could enable significant fishing. The results of this study show these solutions could enable the feeding of everyone given minimal preparation, and this preparation should be a high priority now.

Can't rebuild.

John **Jacobi 17**. [Leads an environmentalist research institute and collective, citing Fred Hoyle, British astronomer, formulated the theory of stellar nucleosynthesis, coined the term "big bang," recipient of the Gold Medal of the Royal Astronomical Society, professor at the Institute of Astronomy, Cambridge University. 05-27-17. "Industrial Civilization Could Not Be Rebuilt." The Wild Will Project. doa: 11/20/24 [https://www.wildwill.net/blog/2017/05/27/industrial-civilization-not-rebuilt/] winky

A suggestion, for the sake of thought: If industrial civilization collapsed, it **probably could not be rebuilt**. **Civilization** would exist again, of course, but **industry** appears to be a **one-time experiment**. The astronomist **Fred Hoyle**, exaggerating slightly, writes: It has often been said that, if the human species fails to make a go of it here on Earth, some other species will take over the running. In the sense of developing high intelligence this is not correct. We have, or soon will have, exhausted the **necessary physical prerequisites** so far as this planet is concerned. With coal gone, oil gone, high-grade metallic ores gone, **no species however competent can make the long climb from primitive conditions to high-level technology**. This is a **one-shot affair**. If we fail, **this planetary system fails so far as intelligence is concerned**. The same will be true of other planetary systems. On each of them there will be one chance, and **one chance only**. Hoyle overstates all the limits we actually have to worry about, but there are enough to affirm his belief that industry is a "one-shot affair." In other words, if industry collapsed then no matter how quickly scientific knowledge allows societies to progress, technical development will hit a wall because the builders will not have the needed **materials**. For example, much of the world's land is not arable, and some of the land in use today is only productive because of industrial technics developed during the agricultural revolution in the 60s, technics heavily dependent on **oil**. Without the systems that sustain industrial agriculture much current farm land could not be farmed; agricultural civilizations cannot exist there, at least until the **soil replenishes**, **if** it replenishes. And some resources required for industrial progress, like **coal**, simply **are not feasibly accessible anymore**. Tainter writes: . . . major jumps in population, at around A.D. 1300, 1600, and in the late eighteenth century, each led to intensification in agriculture and industry. As the land in the late Middle Ages was increasingly deforested to provide fuel and agricultural space for a growing population, basic heating, cooking, and manufacturing needs could no longer be met by burning wood. A shift to reliance on coal began, gradually and with apparent reluctance. Coal was definitely a fuel source of secondary desirability, being more costly to obtain and distribute than wood, as well as being dirty and polluting. Coal was more restricted in its **spatial** distribution than wood, so that a whole **new, costly distribution system had to be developed**. Mining of coal from the ground was more costly than obtaining a quantity of wood equivalent in heating value, and became even more costly as the 54 most accessible reserves of this fuel were depleted. Mines had to be sunk ever deeper, until groundwater flooding became a serious problem. Today, most easily accessible natural coal reserves are **completely depleted**. Thus, societies in the **wake of** our imagined **collapse** would **not be able to develop fast enough** to reach the underground coal. As a result of these limits, rebuilding industry would take **at least thousands of years** — it took **10,000** years the **first time** around. By the time a civilization reached the point where it could do something about **industrial** scientific knowledge it probably **would not have the knowledge anymore**. It would have to develop its sciences and technologies on its **own**, resulting in patterns of development that would probably look similar to historical patterns. Technology today depends on levels of complexity that must proceed in chronological stages. Solar panels, for example, rely on transportation infrastructure, mining, and a regulated division of labor. And historically the process of developing into a global civilization includes numerous instances of technical regression. The natives of Tasmania, for example, went from a maritime society to one that didn't fish, build boats, or make bows and arrows. Rebuilding civilization

would also be a **bad idea**. Most, who are exploited by rather than benefit from industry, would probably not view a rebuilding project as **desirable**. Even today, though citizens of first-world nations live physically comfortable lives, their lives are sustained by the worse off lives of the rest of the world. "Civilization . . . has operated two ways," Paine writes, "to make one part of society more affluent, and the other more wretched, than would have been the lot of either in a natural state." Consider the case of two societies in New Zealand, the Maori and the Moriori. Both are now believed to have originated out of the same mainland society. Most stayed and became the Maori we know, and some who became the Moriori people settled on the Chatham Islands in the 16th century. Largely due to a chief named Nunuku-whenua, the Moriori had a strict tradition of solving inter-tribal conflict peacefully and advocating a variant of passive resistance; war, cannibalism, and killing were completely outlawed. They also renounced their parent society's agricultural mode of subsistence, relying heavily on hunting and gathering, and they controlled their population growth by castrating some male infants, so their impact on the non-human environment around them was minimal. In the meantime, the Maori continued to live agriculturally and developed into a populated, complex, hierarchical, and violent society. Eventually an Australian seal-hunting ship informed the Maori of the Moriori's existence, and the Maori sailed to the Chathams to explore: . . . over the course of the next few days, they killed hundreds of Moriori, cooked and ate many of the bodies, and enslaved all the others, killing most of them too over the next few years as it suited their whim. A Moriori survivor recalled, "[The Maori] commenced to kill us like sheep . . . [We] were terrified, fled to the bush, concealed ourselves in holes underground, and in any place to escape our enemies. It was of no avail; we were discovered and eaten – men, women, and children indiscriminately." A Maori conqueror explains, "We took possession . . . in accordance with our customs and we caught all the people. Not one escaped. Some ran away from us, these we killed, and others we killed – but what of that? It was in accordance with our custom." Furthermore, we can deduce from the ubiquitous slavery in all the so-called "great civilizations" like Rome or Egypt that any attempt to rebuild a similar civilization will involve slavery. And to rebuild industry, something similar to colonization and the Trans-Atlantic Slave Trade would probably have to occur once again. After all, global chattel slavery enabled the industrial revolution by financing it, extracting resources to be accumulated at sites of production, and exporting products through infrastructure that slavery helped sustain. So, if industrial society collapsed, who would be doing the rebuilding? Not anyone most people like. It is hard to get a man to willingly change his traditional way of life; even harder when his new life is going into mines. And though history demonstrates that acts like those of the Maori or slave traders are not beyond man's will or ability, certainly **most** in industrial society today **would not advocate going through the phases required to reach the industrial stage of development**.

Err negative on impact weighing – their evidence is unwarranted pessimism – updated models.

Rodriguez 20 [Luisa Rodriguez is research fellow at the Forethought Foundation for Global Priorities Research. Previously, she researched nuclear war at Rethink Priorities and as a visiting researcher at the Future of Humanity Institute, "What is the likelihood that civilizational collapse would directly lead to human extinction (within decades)? - EA Forum", 24th Dec 2020, doa: 11/20/24 https://forum.effectivealtruism.org/posts/GsjmufaebreiaivF7/what-is-the-likelihood-that-civilizational-collapse-would#Concrete_example__A_large_nuclear_war_that_causes_a_nuclear_winter//imp]

Case 2: 90% population loss, infrastructure damage, and extreme climate change (e.g. nuclear war that caused nuclear winter) In a scenario in which a catastrophe causes the deaths of 90% of the population (800 million survivors), major infrastructure damage, and climate change — for example, a severe, global nuclear war that caused a nuclear winter – I believe the question of whether humans would be able to meet their basic needs becomes more difficult.[14] The questions I consider for this scenario are: What is the likelihood that survivors are able to continue to survive using traditional forms of agriculture, given a catastrophe that causes severe infrastructure damage and climate change? What is the likelihood that radiation causes extinction? What is the likelihood that humanity would survive in the event of conflict immediately following the catastrophe? What is the likelihood that survivors are able to continue to survive using traditional forms of agriculture? Time spent on this section: 2–3 hours Types of sources: Academic literature, non-academic reports, and expert interviews Expert judgment: Several experts, including ALLFED director David Denkenberger, have affirmed this conclusion — they do not expect humanity to dip below the minimum viable population even in relatively extreme sun-blocking scenarios. Literature review: The nature of all of the catastrophes we know of that would cause extreme global cooling (e.g. nuclear winter, asteroid impacts) **would have unevenly distributed impacts** — causing extreme global cooling in some parts of the world, but more moderate cooling in others. For example, in the case of a nuclear war between the US and Russia, nuclear winter models suggest that the most **severe climate effects would be limited** to the Northern Hemisphere, where temperatures would fall by 10–30 degrees C. But in the Southern Hemisphere, and especially at the equator, those effects would be much less severe: between 5–10 degrees Celsius. With heterogeneous impacts like this, it's likely that agriculture would still be possible in some regions – especially in New Zealand and Australia, and possibly in South America and Central Africa.[15] To be clear, I'm describing a very grim scenario, in which basically everyone in the Northern Hemisphere — and in many parts of the Southern Hemisphere — would be unable to grow food using standard agricultural techniques. Given this, I expect there would be mass starvation and violent competition and conflict until a new equilibrium was reached, one where the remaining survivors didn't exceed the Earth's carrying capacity. While I expect this would be a truly terrible period of widespread suffering, I believe this equilibrium would be reached long before the population got anywhere near the minimum viable population. My best guess is the population would fall to hundreds of thousands to tens of millions, but not much lower. While I haven't looked into this much, I feel fairly convinced that hundreds of thousands or **millions** of people **could survive** using traditional approaches to agriculture in parts of the world with more moderate climate effects (and basic mitigation strategies, like switching to crop types that are more resilient to temperature and

precipitation fluctuations). And as with Case 1, at least some of the survivors in a Case 2 scenario would probably be able to survive the immediate aftermath of a catastrophe that caused civilizational collapse by exploiting food and other supplies in stores and larger stockpiles. This would give survivors some buffer time to learn additional skills required to survive once those supplies run out (e.g. fishing) or develop the techniques necessary to produce food using methods that don't rely on climate factors like warm temperatures and regular precipitation. BOTEC: The longer the buffer time, the more likely humanity would be to subsequently survive. But there are a number of different considerations (relative to Case 1) that affect the calculus of just how long such a grace period would be in the context of a catastrophic event like a nuclear war that killed 90% of people and caused a nuclear winter. So I've done a similar exercise to the one above where I try to account for some of those differences. Note: As above, the following BOTEC relies on particularly poor sources, makes a bunch of dubious assumptions (discussed more below), and I'm not confident I've thought of all of the most important supplies. It should be considered very rough. TABLE5 See table note here.[16] Bottom line: I think it's extremely likely that these supplies would last somewhere between around a year and a decade or more. I expect it would be closer to the lower end, given that competition and violence could lead to the depletion of supplies more quickly than if the population were reduced to a smaller number by the catastrophe directly. All this in mind, I think it is very likely that the survivors would be able to learn enough during the grace period to be able to feed and shelter themselves ~indefinitely. What is the likelihood that radiation causes extinction? Time spent on this section: 2–3 hours Types of sources: Academic papers, Wikipedia, and interviews with experts Literature review: In the aftermath of a nuclear war, radioactive fallout from the nuclear detonations would have long-lasting health impacts. In **the most extreme** nuclear war **scenario**s considered by academics (a nuclear war between the US and Russia and their allies, using 10,000 megatons (MT) of nuclear bombs), approximately 30% of the geographic area in the Northern Hemisphere would have enough fallout to be lethal to any adult in the area (Ehrlich et al., 1983). The current US and Russian nuclear arsenals don't currently have that kind of megatonnage (they currently have closer to 2,500 MT). If we naively assume that radiation scales linearly, we might expect a modern day US-Russia nuclear war to contaminate up to 7.5% of the land area of the Northern Hemisphere. This may not sound like much, but consider that 95% of the world's population lives on just 10% of its land area — meaning that 7.5% of land area could be home to millions or even billions of people. What's more, tens to hundreds of millions more might be exposed to enough radiation to be more susceptible to cancer for the rest of their lives. On top of this, there are currently around 440 civilian nuclear power reactors scattered around the world, and likely tens or hundreds more military reactors. These have fail-safes and automatic shut down measures that are designed to ensure that all of the nuclear material in these reactors would be safely contained in the event of a global catastrophe that meant people stopped attending to them. Concretely, these fail safes make sure that water continues to be circulated around the nuclear fuel to ensure it doesn't get so hot it causes a meltdown — i.e., an event where the nuclear core partially or completely melts, which might allow the nuclear fuel to breach its multiple layers of containment and leak out into the environment. If fuel did reach the environment, the radioactive fallout could spread across continents, creating exposure levels ranging from immediately fatal (in areas ranging from tens to thousands of square kilometers) to non-lethal but causing potential higher rates of cancer and infertility. But some of these fail-safes could plausibly fail during a catastrophe that caused infrastructure damage (or afterward, if any components of the fail system degraded). For example, some nuclear reactors rely on backup generators to power the pumps that keep water circulating in the core of the reactor. If those backup generators eventually all broke down, the reactor might melt down. I currently don't have a good sense of how likely these failures would be. Newer nuclear reactors rely on more robust safety systems, with parts that wouldn't break down as easily. And all nuclear reactor safety systems are designed to account for infrastructure damage caused by earthquakes and other physical shocks. But in a large-scale nuclear war, it seems very plausible that at least some nuclear reactors would melt down. My best guess is that this wouldn't happen at a large scale, but even if it did, some areas would likely be far enough away from reactors to be spared the radioactive contamination. For example, Australia has just one nuclear reactor. Even if that reactor were to melt down, much of Australia would likely remain uncontaminated (Australia is just under 3 million square miles, and the Chernobyl meltdown is estimated to have contaminated under 60,000 square miles; and only a much smaller fraction of that area was sufficiently contaminated as to be lethal to humans). Bottom line: While radioactive fallout from nuclear detonations and power plant meltdowns would increase the death toll in the years following the collapse, I expect it **wouldn't be** widespread enough to be immediately **fatal to everyone**, nor would it cause fertility rates or life expectancy to decrease enough to threaten extinction. And at the very least, **some** areas **are sufficiently far away as to be** relatively **safe** from radioactive fallout. What is the likelihood that humanity would survive in the event of conflict immediately following the catastrophe? Time spent on this section: 1–2 hours Types of sources: Academic literature, expert interviews, and speculation Historical base rate: In Case 2, it seems slightly more plausible to me that violence would lead to human extinction than in Case 1, but still fairly unlikely. I don't think human extinction could be caused by a conflict fought with conventional weapons; **there would** just **be too many survivors (~800 million)** to be killed in conventional warfare (compare this to WWI and WWII, during which ~20 million and ~75 million people were killed, respectively). Weapons of mass destruction: My best guess is that the only way violence in the wake of a Case 2 civilizational collapse could directly lead to human extinction is if one group of **survivors** had access to and deployed weapons of mass destruction. This seems unlikely to me, first because it seems hard to imagine a group of survivors incapable of recovering critical infrastructure — and barely capable of meeting even their basic needs — would be able to successfully deploy weapons of mass destruction (though I'm not very confident about this). Second, it's hard to imagine a scenario where the use of weapons of mass destruction kills millions of survivors, spread all over the world, without modern technologies like transportation. For example, with potentially many survivor groups, it seems hard to imagine how nuclear detonations would kill ~everyone despite the fact that the groups would likely be spread out all over the world, potentially in small bands that can't each be individually targeted. Similarly, it's hard to imagine how a pathogen could spread ~everywhere when survivors would likely have greatly reduced mobility (the latter isn't obviously impossible, but it at least seems exceedingly difficult to me). There's one counterargument I find somewhat persuasive, which is that it seems possible that all of the survivors might be confined to a relatively small area (for example, if only a small fraction of the Earth's land area is habitable), making them more vulnerable to a single, large attack. If this were the case, it's easier for me to imagine that the use of weapons of mass destruction could kill all of the remaining survivors. This would presumably mean the aggressors would be killing themselves, which makes it seem even less likely to me. But we've seen humans come dangerously close to threatening their own survival before, often because human aggressors aren't always good at predicting how cascading effects could threaten their survival as well. A random example to make this concrete: If all of the survivors of a nuclear war were confined to Australia, which might be less impacted by a nuclear winter, one group might choose to use nuclear weapons against another group, not realizing that the radioactive fallout or further climate change could make Australia uninhabitable, even for them. Bottom line: I expect the survivors in Case 2 would not deploy weapons of mass destruction against their competitors, as it would likely pose a pretty big risk to the aggressor as well as the target. But I'm uncertain about this — humans have come close to making similarly self-destructive choices before. Thankfully, even if one group did use weapons of mass destruction against their competitors, I still think it's very unlikely that their use would cause human extinction. This is because except in a few very specific and very strange scenarios, I expect the survivors would be too geographically distributed and disconnected to be wiped out by a single act of aggression. I therefore expect the result would be a much higher death toll, but not extinction. Concrete example: A large nuclear war that causes a nuclear winter So what, concretely, do I think would happen in the event of a catastrophe like a nuclear war that led to the death of 90% of the population, and caused severe infrastructure damage and significant global cooling? I expect that, in addition to the billions of people killed in the initial catastrophe, hundreds of millions or more would likely die in the famines and violent competition that followed. But my best guess is that hundreds of thousands to hundreds of millions of the survivors of the initial catastrophe would survive this violent period. I think it's extremely likely these survivors would be able to support themselves using leftover food stocks and supplies, before eventually working out how to feed themselves through traditional agriculture and fishing and/or modified agriculture (using methods that don't rely on climate factors like

warm temperatures and regular precipitation). **All of the catastrophes** we know of **that would lead to extreme cooling** would only do so **for** 1**–10 years, and agriculture would become possible again once the climate began to return to normal**. At that point, it seems even more likely that the surviving humans would be able to meet their own basic needs by returning to traditional forms of agriculture. My key uncertainties are around whether I'm putting too much weight on the idea that humans would figure out how to subsist without traditional agriculture just because it's technically possible, and whether conflict could lead to extinction through channels I haven't foreseen. Another toy calculation suggests that these **uncertainties** probably **aren't troubling enough to change my bottom line**. Note: I again assume each group's fate is independent of the fates of other groups. I actually think this is a pretty reasonable assumption in this case. I expect that the **survivors** of a catastrophe like a severe nuclear war **would end up somewhat spread out** (at least across the Southern Hemisphere), as doing so would create less competition for resources within a smaller area (I discuss this more later). The farther apart the surviving groups are, the less likely they are to be affected by the same shocks (natural disasters, disease outbreaks, conflict). Additionally, in the event of a catastrophe like a nuclear war, transportation, communication, and other technologies that facilitate contact between geographically distributed groups would be enormously limited. This would further limit the extent to which each group's fate ended up relating to another's. There would be other sources of variation between groups that made their fates less correlated: Some groups might be made up mostly of farmers, while others will be made up of lawyers, some groups will tend toward cooperation, while others toward conflict, plus pure randomness (e.g. some groups might have a high proportion of survivors with genetic immunity to a particular disease). But there are also factors that point in the other direction — factors that suggest the surviving groups would be at least somewhat correlated. For example, nuclear winter climate conditions, while nonuniform, would nonetheless impact all surviving groups. Similarly, more severe natural disasters might affect large regions, meaning that at least all of the survivor groups at the regional level might end up experiencing very similar challenges to survival simultaneously. Likewise, there might be things about "human nature" that would be shared amongst all survivors. For example, it's possible that all of the survivors, having witnessed the initial catastrophe, would have similar psychological experiences — like shock, stress, and social distrust, among others — that would make it more difficult to survive and cooperate. As above, the higher the true correlation between survivor groups, the more my toy calculations will cause me to underestimate the probability that all of the survivor groups would be wiped out. TABLE6 With 800 million survivors, the degree of pessimism you have to have about their ability to survive to end up believing that no groups would survive indefinitely is actually kind of extreme. The exact beliefs you'd have to have would depend on whether survivors were concentrated into a few big groups, or distributed in many smaller ones. Specifically: Even if you thought any given group of 100, 1,000, or 10,000 survivors had a 99% chance of being wiped out, it would still be virtually guaranteed that at least one group would survive. If you thought there was a 99% chance that any one of 800 groups of 100,000 people would be wiped out, there would still only be a 1 in 3,000 chance of extinction. The probability of extinction is higher (45%) if you believe that larger groups of 10 million would also have a 99% chance of being wiped out. But, again, to hold that view, you'd have to think that out of a group of 10 million people (again, bigger than the largest US city), not even a few hundred of those people would overcome the obstacles of the post-collapse environment (how to fish, how to farm despite global cooling, avoiding being killed by a hurricane or drought). I do not find this view very plausible. Similarly, the probability of extinction is very high indeed if you think that any given group of 100 million survivors has a 99% chance of being wiped out. Again, to believe extinction risk was that high, you'd have to think that there would be a 99% chance that none of the 100 million people would work out how to survive (for reference, only 14 countries have a population of 100 million or higher). Given all of this, my subjective judgment is that **it's very unlikely that this scenario would more or less directly lead to human extinction.**

Tech outweighs.

Di **Minardi 20**,Communications Officer II Member Of: Ivan Allen Dean's Office, School of Economics, School of History and Sociology, 10-15-20, "The grim fate that could be 'worse than extinction'," doa: 11/20/24 [https://www.bbc.com/future/article/20201014-totalitarian-world-in-chains-artificial-intelligence] winky

What would totalitarian governments of the past have looked like if they were never defeated? The Nazis operated with 20th Century technology and it still took a world war to stop them. How much more powerful – and permanent – could the Nazis have been if they had beat the US to the atomic bomb? Controlling the most advanced technology of the time could have solidified Nazi power and changed the course of history.

When we think of **existential risks**, events like **nuclear war** or asteroid impacts often come to mind. Yet there's one future threat that is **less well known** – and while it doesn't involve the extinction of our species, it could be **just as bad**. It's called the "world in chains" scenario, where, like the preceding thought experiment, a global totalitarian government uses a novel technology to lock a majority of the world into **perpetual suffering**. If it sounds grim, you'd be right. But is it likely? Researchers and philosophers are beginning to ponder how it might come about – and, more importantly, what we can do to avoid it. Existential risks (**x-risks**) are disastrous because they lock humanity into a **single fate**, like the permanent collapse of civilisation or the extinction of our species. These catastrophes can have natural causes, like an asteroid impact or a supervolcano, or be human-made from sources like nuclear war or climate change. Allowing one to happen would be "an abject end to the human story" and would let down the hundreds of generations that came before us, says Haydn Belfield, academic project manager at the Centre for the Study of Existential Risk at the University of Cambridge. Toby Ord, a senior research fellow at the Future of Humanity Institute (FHI) at Oxford University, believes that the odds of an existential catastrophe

happening this century from natural causes are less than one in 2,000, because humans have survived for 2,000 centuries without one. However, when he adds the probability of human-made disasters, Ord believes the chances increase to a startling one in six. He refers to this century as "the precipice" because the risk of losing our future has never been so high. Researchers at the Center on Long-Term Risk, a non-profit research institute in London, have expanded upon x-risks with the even-more-chilling prospect of suffering risks. These "**s-**risks" are defined as "suffering on an **astronomical scale**, vastly exceeding all suffering that has existed on Earth so far." In these scenarios, life **continues** for billions of people, but the quality is **so low** and the **outlook so bleak** that dying out would be **preferable**. In short: a future with **negative value** is worse than one **with no value at all**. This is where the "world in chains" scenario comes in. If **a malevolent group** or government suddenly gained **world-dominating power through technology**, and there was nothing to stand in its way, it could lead to an extended period of **abject suffering and subjugation**. A 2017 report on existential risks from the Global Priorities Project, in conjunction with FHI and the Ministry for Foreign Affairs of Finland, warned that "a long future under a particularly brutal global totalitarian state could **arguably be worse than complete extinction**". Singleton hypothesis Though global totalitarianism is still a niche topic of study, researchers in the field of existential risk are increasingly turning their attention to its most likely cause: **artificial intelligence**. In his "singleton hypothesis", Nick Bostrom, director at Oxford's FHI, has explained how a global government could form with AI or other powerful technologies – and why it might be **impossible to overthrow**. He writes that a world with "a single decision-making agency at the highest level" could occur if that agency "obtains a decisive lead through a **technological breakthrough** in artificial intelligence or molecular nanotechnology". Once in charge, it would control advances in technology that prevent **internal challenges**, like **surveillance** or autonomous weapons, and, with this monopoly, remain **perpetually stable**. If the singleton is totalitarian, life would be **bleak**. Even in the countries with the strictest regimes, news leaks in and out from other countries and people can escape. A global totalitarian rule would eliminate even these small seeds of hope. To be worse than extinction, "that would mean we feel **absolutely no freedom, no privacy, no hope of escaping, no agency to control our lives at all**", says Tucker Davey, a writer at the Future of Life Institute in Massachusetts, which focuses on existential risk research. "In totalitarian regimes of the past, [there was] so much paranoia and psychological suffering because you just have no idea if you're going to get **killed** for saying the wrong thing," he continues. "And now imagine that **there's not even a question**, every single thing you say is being reported and being analysed." "We may not yet have the technologies to do this," Ord said in a recent interview, "but it looks like the kinds of technologies we're developing make that easier and easier. And it seems plausible that this may become possible at some time in the next 100 years."

No pandemic extinction

1] Immunity

Adalja-16 (Amesh Adalja, 6-17-2016, "Why Hasn't Disease Wiped out the Human Race?," Atlantic, https://www.theatlantic.com/health/archive/2016/06/infectious-diseases-extinction/487514/ //AB)

Beyond those three, **every other known disease falls short of what seems required to wipe out humans**—which is, of course, why we're still here. And it's not that diseases are ineffective. On the contrary, **diseases' failure** to knock us out **is a testament to** just **how resilient humans are**. Part of our evolutionary heritage is **our immune system**, one of the most complex on the planet, **even without the benefit of vaccines** or the helping hand of antimicrobial drugs. This system, when viewed at a species level, can adapt to almost any enemy imaginable. **Coupled to genetic variations** amongst humans—which open up the possibility for a range of advantages, from imperviousness to infection to a tendency for mild symptoms—this adaptability **ensures that** almost **any** infectious **disease** onslaught **will leave a large proportion of the population alive to rebuild**, in contrast to the fictional Hollywood versions.

Reproduction

2] Sandberg-14 (Anders Sandberg, 5-29-2014, "The five biggest threats to human existence," Phys, https://phys.org/news/2014-05-biggest-threats-human.html//AB)

Natural pandemics have killed more people than wars. However, **natural pandemics are unlikely to be existential** threats: **there are** usually some **people resistant to the pathogen**, **and** the **offspring** of survivors **would be more**

resistant. **Evolution** also **does not favor parasites** that wipe out their hosts, **which is why syphilis went from a** virulent **killer to a chronic disease** as it spread in Europe.

3[Resilience

Farquhar-17 (Sebastian Farquhar, 01-23-2017, "Existential Risk: Diplomacy and Governance," Global Priorities Project, https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf//AB)

For most of human history, natural pandemics have posed the greatest risk of mass global fatalities.37 However, there are **some reasons** to believe that **natural pandemics are very unlikely to cause** human **extinction**. Analysis of the International Union for Conservation of Nature (IUCN) red list database has shown that **of the 833 recorded plant and animal** species **extinctions** known to have occurred **since 1500**, **less than 4%** (31 species) **were ascribed to infectious disease**.38 **None of the mammals and amphibians** on this list **were globally dispersed**, and other factors aside from infectious disease also contributed to their extinction. It therefore seems that **our own species**, **which is very numerous**, globally **dispersed**, **and capable of** a **rational response** to problems, **is** very **unlikely to be killed off by a natural pandemic**. One **underlying explanation** for this is that **highly lethal pathogens can kill their hosts before they** have a chance to **spread**, **so** there is a **selective pressure** for **pathogens not to be highly lethal**. Therefore, **pathogens are likely to co-evolve with their hosts rather than kill** all possible hosts.39

Rebuttal:

No ev read – marked the doc