

Opinion | We Are Clueless About What's Coming

Spencer A. Klavan, Sophia Muniz

Advertisement

[SKIP ADVERTISEMENT](#)

interesting times

Nick Bostrom is no longer worried about the apocalypse. Now he's worried about utopia.

Oct. 1, 2026

By Spencer A. Klavan

Produced by Sophia Muniz

Video

Silicon Valley's Favorite Prophet Has a New Warning

Nick Bostrom is no longer worried about the apocalypse. Now he's worried about utopia.CreditCredit...The New York Times

Nick Bostrom rose to prominence for predicting artificial intelligence doom more than a decade ago — garnering support from figures like Elon Musk, Sam Altman and Bill Gates. But now he's retreating from those warnings, focusing his attention on a more complex question: What if A.I. succeeds beyond humanity's wildest imaginations? Will we be bored out of our minds? Are we too infantile to understand what's coming?

Guest host Spencer Klavan, the associate editor at the Claremont Review of Books, presses Bostrom on the potential uses and abuses of superintelligence, rationalist ethics and what makes us human.

Silicon Valley's Favorite Prophet Has a New Warning

Nick Bostrom is no longer worried about the apocalypse. Now he's worried about utopia.

Listen · 50:45 min

Below is an edited transcript of an episode of "Interesting Times." We recommend listening to it in its original form for the full effect. You can do so using the player above or on the [NYTimes app](#), [Apple](#), [Spotify](#), [Amazon Music](#), [YouTube](#), [iHeartRadio](#) or wherever you get your podcasts.

Spencer Klavan: Nick Bostrom, welcome to "Interesting Times."

Nick Bostrom: Hi, Spencer. Good to be with you.

Klavan: I wanted to start by getting your assessment of where you think things stand. I think the technical term of art I want to invoke here is P (doom), the probability that everything goes terribly wrong.

A researcher at Anthropic recently placed this at around 10 percent or maybe even above. Is that about where you are? Do you have a different assessment? At this point, what do you think is the likelihood of catastrophe?

Bostrom: I don't think my outlook has changed fundamentally. I see it more that the world has changed.

When I started working on the first book, "Superintelligence," in 2008, at that point, the world was in a very different place. People were dismissing this whole thing as science fiction. And nobody, save maybe a handful of individuals, were thinking about the challenges that would arise if we one day developed really powerful forms of machine intelligence.

At that time, it seemed most useful for me to try to draw attention to the potential pitfalls on the path ahead so that we could use the intervening years to prepare ourselves — to do our homework, basically, like to do the research on alignment, etc.

Now, of course the discourse has changed a lot, and now it's very mainstream to think that superintelligence is not just a science fiction-like thing, but a real possibility.

So it seemed less useful for me to just keep ping-ponging the same string. There was this other side of the coin, which was already always there in my mind — like the upside and the downside — and to just take a glance at what would happen if things go well. Like, if we take the best case scenario, that also has philosophically problematic aspects to it.

Klavan: I'd like to get into some of the specifics of what those scenarios might be in a bit, but these numbers do get very specific. I've invoked one Anthropic researcher, but I've seen a lot of people saying P (doom) is 70 percent or 10 percent.

Can you walk us through the process of assigning a number to a probability like that? Is there a math behind it? How does that work?

Bostrom: No, there isn't. The number itself comes out of a black box and different people — Those are meant to be expressions of a subjective degree of uncertainty and confidence.

It's not possible to calculate the number based on some kind of set of agreed assumptions, which is reflected in the fact that different people come up with very different numbers.

My view is that there will be real risks associated with this transition to the machine superintelligence era, including existential risks. We should take those seriously.

If you want a more qualitative assessment, I would say that I'm a fretful optimist.

Klavan: As in, you think that there's a real chance things go quite pear-shaped, but you're hoping, you believe it's not forever concluded.

Bostrom: Yeah, I'm really excited about the upside. I'm hopeful we could get there, but I worry that we might not.

Klavan: Let's stay in our anxieties for just a second and wallow in them together. Can you put some more meat on the bones of these potential doomsday scenarios? What are some of the things that could go wrong? What would that look like?

Bostrom: I think there are several challenges that we would need to meet in order to get a good outcome here. One is the alignment problem. This is the technical difficulty of developing scalable methods of A.I. control. Basically, ways to ensure that we can continue to steer these systems and have them do the things we intended for them to do, rather than something different, even as these systems become ultimately arbitrarily capable. That's one challenge.

I think a different class of challenge is what you might call the governance problem. Assuming we manage to control this technology, there is then the question of what ends will we humans put it to. Sometimes we use our powers for ill purposes — to wage war, to oppress each other, maybe to concentrate power — to do all kinds of nefarious things.

So the second challenge is: Can we at least do a reasonably decent job of shifting the use cases predominantly toward the positive?

Klavan: Some of these are touching on the scenarios that you were contemplating back in 2014 and thereabouts — the question of staving off or keeping people from inventing disaster technology. I'm thinking especially of your paper on the "vulnerable world" hypothesis, the idea that there exists some tech somewhere that, if invented, will or could instantly cause disaster.

In that paper, you raise a bunch of things that might be necessary to keep this from happening. You talk about the possibility of mass surveillance, one world government — all kinds of things that would be, I think, pretty difficult and daunting and terrifying to contemplate now, but you say maybe this will start to appear necessary.

Are you still contemplating measures of that kind? Is that still a possibility?

Bostrom: Well, I should clarify that the paper doesn't advocate any of those things.

Klavan: No.

Bostrom: It considers this possibility, if we are unlucky. And then I consider if that were the case, then what would have to be true of a civilization that invents that technology in order for it not to be devastated?

Depending on the nature of that technology, if it's one that sort of incentivizes first strikes among major powers, then it might require some kind of global coordination. If the problem with the technology is that it makes it too easy for random individuals to bring down civilizations, some sort of super weapons of mass destruction that you could concoct in your own kitchen using ordinary materials, then it looks like you would need to do this very detailed surveillance mesh to prevent people from doing that.

But the world is not guaranteed to be pegged up such that we will have an easy time. Anyways, so that's that paper. It just tries to create a framework for thinking what the different possible situations could be that we might find ourselves in.

Klavan: OK. Let's shift our focus now and start thinking about what happens if things go right — a cheerier subject, maybe. You've been writing it since at least 2024.

Your book "Deep Utopia" is about the scenarios we can hope for, the optimal setup. Can you lay that out for us? What does it look like if this goes well?

Bostrom: If we develop this strong form of superintelligence, then I think what that means is that across the board, across all other technological and economic fields as well, we would have extremely rapid advance because these superintelligent minds would then be doing the science and the research and the economic activity, but at digital speeds.

I think you would have a kind of rapid galloping toward perfect virtual reality or a cure for aging or space colonies or all kinds of almost science fiction-like technologies that don't violate any laws of nature. They're just really hard.

So the question then becomes: In such a condition of technological maturity, what would a good human life look like? What would a future path for us look like under these, in some sense, ideal circumstances?

I have this concept of a solved world, which is the idea of technological maturity plus some sort of decent job on the governance things.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning.

Let's imagine everybody gets a slice, everybody is free. What happens then is that a lot of the constraints that currently give shape to our human lives would peel away.

You could have the A.I.s and the robots doing basically all economically productive tasks. You then have a leisure society where there is no longer any need for humans to work in order to get a paycheck.

If that was all, then I think it would obviously require some cultural transformation and adaptation. But ultimately I think we could easily imagine living in such a world.

There are already various groups of people who don't work for a living that you could look at for inspiration. Young children don't work for a living. They seem to have a good life. They spend time with friends, playing games, inventing things.

Klavan: It depends if you aspire to live like children.

Bostrom: Then we have retired people. We have a confounder there, that many retired people have poor health and stuff, but if you imagine somebody who has enough money and perfect health, they often have great lives. They go around traveling, spend time with their grandchildren. They have hobbies. They watch movies. They golf. Whatever it is, it's not necessarily a dystopian existence.

Klavan: Sounds great.

Bostrom: Then you have various other people who maybe have a lot of independent wealth. They don't need to work for a living. I think there would be changes.

Like, the education system right now is configured to take in children and then output productive workers with quality grading attached —

Klavan: This is one of your rants in your book, "Deep Utopia."

Bostrom: Yes, it is my rant. I do have sympathy because right now, the world does need a lot of productive workers, so in some sense we have no option. But if we no longer needed productive workers, then it would be perverse to continue to think of education as just a preparation for entering into the labor force and to get a desk job at an office.

Maybe then have an education geared more toward preparing people to live a great life, cultivate the art of conversation, hobbies, spirituality, physical wellness, arts —

Klavan: Now you're speaking my language. I'm a college professor. I teach Greek literature.

Bostrom: Oh, nice.

Klavan: You don't have to sell me on an education for a life of leisure.

But can we get into what happens then, when we peel back that layer of the onion?

Bostrom: So then you think a little bit harder about this and you realize that it's not just our economic efforts that would become otiose in a solved world, but also a lot of other instrumentally motivated effort.

Somebody who wants to be physically fit has to go to the gym themselves. They can't hire an assistant to lift the weights on their behalf —

Klavan: Much as I would like to. That would be great.

Bostrom: Yeah. Maybe another person enjoys decorating their house. They want to pick out the curtains, the pillows, go through the catalogs, and then at the end, they can get it in just a way that is right for them, that they really feel reflects their own preferences.

But at technological maturity, you would have recommender systems that would know your preferences so well that it could just select the curtains and the cushions and the rest, and do a better job than you would if you went through the trouble to do it yourself. You could still do it yourself, but it might seem pointless.

You can walk through the different things that people fill their time with when they don't have to work and it does seem there would be this question of: Is there still a point to doing this if there was a shortcut, a button you could press that would result in the same outcome?

Klavan: Isn't the endpoint of all this that you press a button and you feel infinite pleasure, or you're suddenly infinitely wise? That you reconfigure your environment, your mind? I've read you talking about this.

Bostrom: Yes, you have this plasticity. The world becomes plastic in the sense of malleable to our desires, including ourselves. There's a lot of things people do now for the sake of getting a certain mental state.

Klavan: Yeah.

Bostrom: Maybe the runner wants the endorphin boost at the end of it, or we engage in some activity because it gives us joy. But at technological maturity, you could directly induce the endorphin or the joy.

So it does raise this more profound philosophical question of what ultimately has meaning. What is worth doing for its own sake, not because it is practically useful to achieve something else?

Klavan: What exactly does it look like to get there? Let's walk step-by-step from here in our weak, fragile human bodies that live and die, to deep utopia.

You open up a bunch of possibilities of how I might get from here to there, including genetic modification and eugenics. I'm not saying that these are things you've said we should do.

Bostrom: No. I think that's important to clarify. There is this chapter in "Superintelligence," the earlier book, where I consider different possible paths that might result in superintelligence.

One is whole brain emulation, where instead of building these completely artificial synthetic minds, you would try to do something that is very closely based on how the human brain operates, like maybe detailed brain scans and then implementing neural networks.

Sort of copy pasted from this, that would be a second route to machine intelligence. Then you could imagine forms of improvements in our collective epistemology institutions and tools that allow humans to solve much more difficult problems together. So that would be like a kind of collective superintelligence.

Then you could imagine maybe some smart drugs, nootropics that would improve our memory or concentration. Or you could imagine some form of genetic enhancements.

But I think the story began a long time ago, when somebody developed a better way to chip a piece of flint to create a better ax, and then better ways of planting seeds, and then all kinds of inventions — double-entry bookkeeping — all the way up to the current era.

Maybe we're getting close to the point where we can automate a lot of this process of inventing that could then speed things up radically. That would be the most plausible path if we do end up at technological maturity.

In our lifetime, certainly, it would probably go through the route of machine superintelligence.

Klavan: That was going to be my next question. Do you, Nick Bostrom, want this? Do you want a world of radically transformable minds in an electric universe?

Bostrom: Before answering that question, I guess it's worth just reflecting for a second on what it would mean if we said no, even though we had the ability.

It's kind of the embedded telos in so much of the modern world. In science, we're trying to figure out more and more about how the world works. In technology, we try to increase our ability to accomplish goals more efficiently.

The economy tries to invent new processes and ways of organizing so that we achieve our goals with fewer inputs. So if one rejected the endpoint, it would then call into question: If you backchain from that, thinking it's perverse to be working so hard to achieve a goal such that if we achieve the goal, it would be very bad for us.

Klavan: Well, there's another account of that you could give, which is if you follow technology's goals unfettered or unmoderated by any other goals and aspirations, you end up in a pretty dark place. But that means maybe technology's not the highest good, or there are other goals that we might seek?

Bostrom: Sure. I mean, technology's a means rather than a goal. But nevertheless, those means would remove more and more constraints in our ability to achieve outcomes.

Klavan: So having no constraints is the goal? And then there would be ——

Bostrom: In some sense, it's implicit in technology. But then it does turn out that if you actually follow through that logic, it becomes a lot less obvious how careful we actually are about achieving that goal.

Klavan: I guess again that's what I'm asking.

Bostrom: It's like a dog chasing a car.

Klavan: [Chuckles.] Yeah, once you catch the car ...

Bostrom: We're putting out all this effort. We hadn't really thought about what would happen if we succeeded.

Klavan: So maybe technology's goods and technology's goals aren't the only goods or goals that are choice worthy, that we should want.

Bostrom: Well, I think that is a given. It's not determining all the goals that we ultimately want. But you can then think about what would happen, given all the other goals that we might have, if the technology part runs its course, and you end up in this solved world.

Then the question is: How suitable is a solved world for the realization of all these other goals that we might have? That's the question that "Deep Utopia," the book, is wrestling with.

I think that there is at least an initial level of repugnance. Once you stare at this, at least a part of you is apt to recoil.

Klavan: I feel this. Yes. I will put my cards on the table.

Bostrom: So the question is: How much can be rescued, of what we care about?

If you push through, you can ultimately come out on the other side of that and think that it could be an extremely valuable situation.

But let's first think which of our values could you have in a solved world? What is possible? And we can sort of walk through a few different things.

Klavan: I'm still a bit stuck on, for now, the journey from here to there. As we contemplate what to do in

this moment, part of what we're contemplating is whether we want to let the technology run and spool all the way to a world of total plasticity.

I began by asking: Is that a positive outcome for you? Does that look like utopia, to be able to change your environment at a whim? And I don't know yet what your answer is.

Bostrom: We could decide to retain various hamster wheels for ourselves to run in. Normally, if somebody creates an engine and it pollutes less and requires less fuel and is cheaper to build, we celebrate them. But you could imagine running that in reverse and sort of applauding people for making our systems less efficient. If some government department can waste more money, you could say that's good because it brings us farther from this telos of maximum efficiency. But I think that's another counterintuitive stance to take.

Klavan: Given that we have arrayed before us these many glittering possibilities, and we've said that what technology tends toward is not necessarily good or the only good, I'd like to know a little bit more about what goods you are aiming at. What are the other things that we should pursue or try to preserve as these developments unfold?

Bostrom: I think it might be useful to start with the simplest, and then sort of build up to the more subtle or difficult or deeper value.

Klavan: OK.

Bostrom: If you start at the very most obvious level, people might care about pleasure and avoiding pain and just having positive subjective experience changing our own mental states. Maybe there are like super drugs without side effects or maybe other ways of rearranging and intervening in our nervous systems, such that if the utopians wanted to actually enjoy every day, they certainly would have the means to do so.

Klavan: OK. But I think most people ——

Bostrom: Want more?

Klavan: ——Don't feel that way.

Bostrom: Yeah.

Klavan: I think you don't feel that way.

Bostrom: I think there is more that I would reach for. I do think it is one important component that is easy to dismiss when you're trying to make some clever philosophical argument or sound profound or something pleasure, ha ha.

But actually, if the future is without that, I just think it looks way bleaker than if it's something where the inhabitants of this are actually having a good time.

You don't have to think of the utopians as being spread out on a mattress like a drug addict on a euphoric high, but they could sort of be admiring all the beauty that is maybe around us already now, in so many ways, but we are oblivious to it or dull.

That already, I think, would somewhat add to the appeal of these utopian lives, beyond the appeal of the sort of hedonistic pleasuremaxxing junkie. But there is more.

Klavan: OK. What else? What else can I have?

Bostrom: So far we were imagining this as a very passive form of existence. You sit there, you take in all of this wondrous beauty. Maybe that's really nice. You enjoy it. But still you might think: Well, what about activity? Don't we want to do things? So you could add that. You could have various forms of

activity, exertion, effort.

Now, here we have the problem we discussed earlier, that there might be a pointlessness to a lot of things. But we can have an artificial purpose. So if somebody sets themselves the goal of getting a little ball into a sequence of 18 holes using the very inconvenient method of hitting it with a club, after you have that goal, then the only way you can achieve it is by you yourself concentrating and taking swings.

So game playing might constitute a larger portion of the lives of utopians. Games are essentially where we set ourselves some arbitrary goal in order to enable the activity of pursuing it, that we might think is intrinsically valuable.

Klavan: At this point, I have to say, you've laid out for me this way of looking at the future of A.I., which seems to suggest we've got this technology, it's coming into being, it may attain all these new capacities really quickly, maybe overnight. And the range of options include, "It's going to nuke the world," all the way up to, in the best case scenario, "It's going to put us into a sort of digital bath," where we are uploaded minds or we're endlessly customizable bodies that feel forms of pleasure, appreciation, interest, all of which are detached from real stakes or requirements except those that we impose, agree to, or imagine.

And I have to say, none of these sound like very good outcomes to me. I don't even want somebody else doing my weightlifting for me, and I'm not sure that this is a utopia that most people would look at and say: Yeah, I recognize something good in that.

I still keep feeling ——

Bostrom: It's not a sales pitch.

Klavan: No, it's not. [Chuckles.]

Bostrom: It's an analysis. You might still be missing a little bit here. So we talked about artificial purpose that allows activity. It would not be, like, a bath. They might go around participating in all kinds of activities and sport and music performances and all kinds of things that are beyond even our current conception of what the rich human life could look like.

Just as our lives are maybe beyond the conception of, if you can imagine, our great ape ancestors. They might have thought: Oh, if we evolve into humans, we could have so many bananas. Now we do have a lot of bananas, but there's more to human life than bananas. There is romantic love and literature and religion and politics and humor and all of these things that they literally couldn't imagine because the ape brain wasn't capable. It would be kind of preposterous, I think, to think that we have somehow maxed out and can now see all possible things that would be valuable.

But I want to squeeze in as well, the question of: Are there real purposes that could survive into a solved world? Things we would have reason to do, not because we set ourselves some arbitrary goal just for the sake of doing stuff, but that are more independent of us.

I think there could be some. For example, suppose you happen to have the value of honoring and valuing certain traditions that you are committed to and feel part of, maybe, and that you want to carry forward. It wouldn't count if you built a robot that sang hymns or enacted the ceremonies at Christmas or whatever.

To the degree that we care about what other people think of us, if what those other people care about is that something be done by you, by your own effort, that would give you a reason to do those things inasmuch as you care about what those other people think. It's not arbitrary.

It's the same thing as if your child draws you a crayon picture for your birthday, and even if a robot might do something nicer or it might be that you could download a nicer picture from the internet in some objective sense, it still means a lot more to you precisely because the child chose to do it and put effort

into it. So that gives the child a reason to put this effort into it. Beyond that, all kinds of spiritual pursuits and a relationship to values.

I have this metaphor — this might appeal to you slightly more than some of the earlier discussions — that right now there are obvious values that are stark and urgent. People starving, people suffering from disease. There's poverty, injustice. These have an immediate claim.

Beyond that, there might also be subtler values that are like the constellations in the night sky. They are actually there during the day as well. It's just that they are blotted out by the intensity of the sun. These are these screaming moral imperatives and urgent practical necessities.

But if you imagine a world where that went away, then I think it would make sense to let our pupils dilate, and we might then see this much richer canopy of values — reasons for doing things that are maybe more aesthetic, more spiritual, calling forth a greater sensibility from ourselves, doing things because actually it would kind of be beautiful to do it this way rather than this other way, even though the other way is in some sense simpler and easier and more direct.

Klavan: OK. This leads me to a question that I had. So you say these aren't sales pitches, they're predictions or analyses.

Bostrom: No, they're not predictions either.

Klavan: OK.

Bostrom: It's an exploration of what happens —

Klavan: What could be.

Bostrom: — If we end up in this situation where these philosophical questions then become very obvious.

Think of it as a philosophical particle accelerator. You can create extreme conditions in an experimental setting to see the basic principle. Similarly for our values, I think we can gain insight into what ultimately is value and what we just associate with value by thinking of this extreme condition of a solved world where a lot of the practical confounders are removed by stipulation.

So it's not a prediction that that's where we will end, but it's a philosophical setup that can sort of force you to confront, at a deeper level, the question of what you ultimately value.

Klavan: But values are precisely what I am not yet clear about from you. I've heard a lot of different possibilities and these thought experiments that we can put into the particle accelerator and the variety of things that we might value as futuristic mega minds. And yet, I don't know which of those things from our current standpoint we should be aiming toward, and which of those things we should be discarding.

We're contemplating, for instance — if we are going to be altering ourselves in this direction — which parts of our humanity are kind of expendable and which parts we should maximize or optimize?

In this context, I think it would be helpful if we can talk about your philosophical outlook a little bit. I know you are associated — you are pretty influential in the effective altruist movement. That's the movement that seems to be defining a lot of these doomsday scenarios that are coming out of Silicon Valley that we've talked about.

Can you define what effective altruism is, what it means, and talk about how your work relates to it or doesn't align with it?

Bostrom: First of all, I wouldn't call myself an effective altruist. I have a lot of friends in those circles, but I think they would maybe say that they try to be rational about doing as much good as possible from

an impersonal point of view.

A lot of traditional philanthropy is some rich person trying to show how rich and well-connected they are to other rich people, and they give to this already very wealthy symphony orchestra in their town because they throw the parties for all the other interesting rich people. Is there maybe a better way that you could do a lot more good with whatever number of dollars you're willing to give?

So it started then as an attempt to do philanthropy more efficiently, particularly third world philanthropy, with bed nets and the deworming medicines like vitamin pills, iodine, etc. And then from there, I think it broadened out to also look at other cause areas where the stakes are more abstract, but could have a disproportionate impact on the future.

My own role in this, I think, is that some of the concepts and ideas I had were maybe helping to widen the scope over which cost-effectiveness prioritization was done, so not just within a particular domain.

Klavan: This was at our shared university home in Oxford.

Bostrom: Yeah, there was this Future of Humanity Institute at Oxford, which I ran for many years. A couple of the people involved there were also pioneers in this effective altruism movement.

And there has long been a great level of interest among many effective altruists in these bigger picture questions for humanity. What are the major risks to human survival, ways we could destroy Earth? Because if you take the stance of trying to do the maximum good impersonally, then it does look like one important dimension is how our actions now affect things over time, including the longer future, not just the immediate effects.

If you care a lot about the future where the most people will live, and they think they are equally morally considerable as the current people who live, then ——

Klavan: Do you agree with that, that the future people outweigh or are as morally considerable as present people?

Bostrom: I think that maybe we have more reason to place extra weight on a lot of current people. It's not as if the future generations have monopolized almost all the efforts. In general they tend to be very shortsighted and selfish. I think some stretch in that direction is probably for the good. But it's possible to take it too far, and then you might become a kind of fanatic.

Like, you can't answer all, the question of life, the universe, and meaning all in one little manifesto, right?

Klavan: No, but this is exactly now in this conversation what I'm asking you to — since we have actions, choices we have to make. This is what I want to hear from you.

Bostrom: Yeah. So I'm more of a pluralist, an ethical pluralist. I think of moral decision-making in terms of this metaphor I have of a moral parliament, where you could imagine different ethical theories. There's the utilitarian party, there are the ontologists, virtue ethicists, ethics of care. There is also a self-interest party — I care about myself and my family particularly. All these different things.

Maybe I have slightly different degrees of credence in these different ones, so they get to send delegates proportional to the probability or the weight that I give them. And then in this imaginary parliament, all these delegates convene, and I imagine them negotiating with one another to then determine a policy that would be what I choose to do.

I think that that would tend to be more robust than trying to guess which is your favorite moral theory and then devote yourself fanatically to pursuing whatever you might think that theory says you should do.

Klavan: OK. But what about the scenario that we're in right now? I'm thinking now again of the vulnerable world hypothesis and the idea of using mass surveillance to stave off some bad technology. At

some point, you have to think about, if we do arrive at that juncture, are we going to accept a freedom tag on our ankle? Or is that in itself an unacceptable curtailment on human freedom?

You've laid out a way that you might have an intellectual parlor conversation, a philosophical dialogue among a variety of different ethicists about that. But I want to know how you, Nick Bostrom, would make these sorts of decisions?

Bostrom: So how would I decide? I think it would depend a lot on the specifics of the situation.

Klavan: OK.

Bostrom: I mean, I would say that there is a propensity for a society to opt for the surveillance option even when they shouldn't, like after 9/11, for example. It was like a few thousand people who died, and yet it then triggered a multitrillion-dollar effort, throwing civil liberties overboard, launching a massive surveillance program, invading countries, and trampling over human rights in some instances.

So I have a lot of sympathy with the idea that the greater risks would be that we jump too readily to say more surveillance or even more global governance rather than the opposite. But that doesn't mean that you can't conceive of situations where maybe that would be the lesser of two evils.

Klavan: Well, it does seem as if when you're gaming out these scenarios many, many hundreds, thousands, millions of years into the future, you start to weigh the concerns of the present against this enormous mass of uncountable millions of humans, post-humans, biohacked minds, digital minds. And suddenly our present scenario can't weigh in the balance.

So our little puny humanity has to accept — your example of curtailments on freedom after 9/11 is a good example — this big bogey comes up on the other end of the future, and then suddenly in our now, we're making our lives miserable. And a lot of these things that you're describing do sound like making our lives miserable now because we've accepted an idea about what might come down the pike.

How is it that these kinds of scenarios help us to think about what we do now, rather than just paralyzing us or sending us into a panic?

Bostrom: Well, there's a lot there. I wrote this paper, maybe back in 2003 or thereabouts, called "Astronomical Waste." It basically says that there are a lot more future generations that could come into existence, so the concerns of the present generation from that ethical point of view would seem to be extremely small and low weight compared to the combined interest of the entire future that is on the line.

Now, the mistake that people make is that they think that I think that we should give basically no consideration to the current generation and focus all our effort on ensuring that this glorious future comes about. But the paper says that that's what follows if you accept certain assumptions, these utilitarian premises, this utilitarian ethic.

Now, I am not myself a utilitarian. And indeed, one argument against being into utilitarianism is precisely that it would have these quite counterintuitive implications, which the paper helped point out. I think there are also some other problems with infinite ethics, etc.

So my actual views are a lot less crazy, let's say, than what people like to believe.

Klavan: I take this point, and I think it's a useful one, that you view your role as a philosopher as opening up a variety of possible scenarios and then gaming them out according to various different ethical approaches. People can then draw their own conclusions, let's say, about whether that looks like a good ethical approach.

At the same time, you're a tremendously influential figure. There are people right now contemplating real world decisions, which we began with, about how to approach A.I. development, whether to slow it down, whether to regulate it — and they are making reference to your work.

I accept that you don't think that's an appropriate use of your work or it's not what you intend your work to be used for, but I come back to the fact that when I asked you what kind of utopia you want to live in and how much of our humanity you're willing to edit away or augment, I didn't have a clear picture of what a good world, a good life looks like for you.

That's, I think, a problem. If we're looking for guidance now about how we should proceed in these major questions, it seems like this way of thinking is actually kind of a choice paralysis rather than a helpful guide.

Bostrom: Yes. I mean, I wish I had more guidance to give.

Klavan: [Chuckles.] OK.

Bostrom: I would say I tend to think not so much in terms of an optimal state to reach, but more in terms of a trajectory, of development from —

Klavan: You have a paper called, "Why I Want to Be a Posthuman When I Grow Up."

Bostrom: Oh, that's an early one, yes.

Klavan: OK. But you no longer feel that way?

Bostrom: Well, no, I do think, but the trajectory here is important. I'm thinking this is all subject to uncertainty. Maybe I would rethink this if I were wiser, etc. But if you imagine a path where instead of decaying and dying after a few short decades, as you have started to accumulate some wisdom and experience, the brain rots and you forget everything and become senile, and then it's all gone.

You could imagine, if you could stabilize the situation, maybe with some longevity therapy or something, give us a little bit more time, maybe alleviate some of the worst forms of suffering that are just too desperate and shouldn't happen, then there would be less urgency. You could then maybe exhale a little bit, have time to take it slow, to do the various things that you can do as a human.

There are people who find great value in basket weaving and others who enjoy sailing boats and somebody who becomes a musician or just studies all these great books, because there are so many of them that even if you're a scholar just studying great books, you still don't really have time.

Because a great book, you don't want to just read through it quickly. You need to live with it for a while to really take the urgency off there.

Klavan: Tell me about it. I'm familiar with the problem.

Bostrom: I could imagine if you had lived like that for some period of time, I don't know whether that is 150 years or 500 years or 2,000 — we've never lived that long, so we don't know — then at some point, maybe, it would start to feel like you have kind of done the human thing. Like you've kind of been there, done that, and maybe it's time to unlock the next level.

So that may be a slight increase in your mental capacities or in your ability to feel deeply, like some further little layer on top that you could then grow into and start to explore these more, I don't know, transhuman ways of being, and new things would maybe appear in view. Just as with the apes evolving into humans, there would be new things that we can do as humans that we think are really beautiful and valuable.

And then slowly over time, maybe the end result after millions of years of this, maybe eventually you do end up as some kind of strange planetary-sized post-human super mind that is kind of connected to other super minds and doing who knows what.

It's hard for us from our current vantage point to evaluate that. I think we are sort of, in a deep sense,

clueless about that.

Klavan: I want to react to what you've just said in the, in the way that we were talking about reacting to your work as a possible future to contemplate, and then I want to let you have the last word.

As you say, thinking about these things provokes reactions in us that reveal our values and our ideals and what we want. And I think part of the thing that makes this deep utopia look almost as horrifying to me as the disaster scenarios we began with is that I actually like humanity.

I think all of the values that we might even select to aim us into the future, like beauty, like justice, like love or truth, these are all features of the kinds of beings that we are, just as our current limitations and pains and mortality are part of our humanity.

And so when you present this scenario in front of me that we could transcend that into something that we can't even imagine, which would have goals and aims and aspirations completely beyond my current conceptualizing, to sort of paraphrase Liz Lemon: I do not want to go to there.

Like, down that road, I do not want to go. And I wonder if you can see why a lot of people might, as you put it, recoil at this kind of prospect.

Bostrom: Yeah, I can see that. But if you consider a young child who eventually grows up, it is true in one sense that the grown-up they become is very different from the child. They have different interests, different thoughts, their personality, their body is quite different. And yet we don't normally think of it as, all things considered, bad for a child to grow up.

I think in some larger sense, we might all still be like children. And I'm hoping that we can continue to develop and carry on some of what is valuable about what we are as a child. Hopefully we can do a better job at preserving that than what the natural maturation from child to adult involves.

So we might remain a child, but also then grow these additional layers on our being that would allow us to become, in some sense, even more of what we are in potential and unfold that.

I just feel like there's this vast space of possible modes of being, of ways of feeling and doing and relating, that it would be hugely presumptuous to imagine that the little janitor's closet that we so far have explored would somehow be the best possible place in this giant cathedral.

I don't think that path is one in which we would be making these choices without external contacts and considerations. Imagine it's not an isolated person shooting off into space in their own space capsule necessarily. It could be a community, a civilization, a larger entity, maybe joined with — you could imagine, if you want to be fanciful — uplifted animals that can start to talk and participate, maybe some of these digital minds.

And then together, feeling our way forward into what might be a beautiful future for us all.

Klavan: I wish you the best on that journey. I won't be there. Nick Bostrom, thank you so much.

Bostrom: Thank you.

Thoughts? Email us at interestingtimes@nytimes.com.

This episode of "Interesting Times" was produced by Sophia Alvarez Boyd, Victoria Chamberlin and Raina Raskin. It was edited by Jordana Hochman. Mixing and engineering by Pat McCusker and Efim Shapiro. Cinematography by Marina King. Video editing by Dani Dillon, Julian Hackney, Arpita Aneja and Steph Khoury. The supervising editor is Jan Kopal. The postproduction manager is Mike Puretz. Original music by Isaac Jones, Sonia Herrero, Pat McCusker and Aman Sahota. Fact-checking by Kate Sinclair, Mary Marge Locker, Julie Beer and Michelle Harris. Audience strategy by Shannon Busta, Kristin Lin and Andrea Betanzos. The executive producer is Jordana Hochman. The director of Opinion

Video is Jonah M. Kessel. The deputy director of Opinion Shows is Alison Bruzek. The director of Opinion Shows is Annie-Rose Strasser. The head of Opinion is Kathleen Kingsbury.

The Times is committed to publishing [a diversity of letters](#) to the editor. We'd like to hear what you think about this or any of our articles. Here are some [tips](#). And here's our email: letters@nytimes.com.

Follow the New York Times Opinion section on [Facebook](#), [Instagram](#), [TikTok](#), [Bluesky](#), [WhatsApp](#) and [Threads](#).

Advertisement

[SKIP ADVERTISEMENT](#)