Rogue A.I., Bioterror, and Other Ways Tech Might Take Us Out

Bryan Walsh

Martin Rees is the 15th Astronomer Royal, an actual senior post in the Royal Households of the United Kingdom. That means, as he writes half-seriously in his new book, *On the Future: Prospects for Humanity*, if the Queen of England wanted her horoscope done, "I'm the one she'd ask." Rees isn't actually an astrologer — he's a cosmologist and astrophysicist, one who in a long and distinguished scientific career has made breakthroughs on black holes and the tantalizing possibility of the multiverse, while finding time to serve as president of the Royal Society and as master of Trinity College at Cambridge University, where he still works today at the age of 76.

But though Rees can't read your future in the stars, he is in the business of prediction. In his 2003 book, *Our Final Century: Will the Human Race Survive the Twenty-First Century?*, Rees raised the alarm over the dangers that emerging technologies like genetic engineering and artificial engineering could pose to the future of the human race. (The book was published in the United States with the title *Our Final Hour*, because, Rees joked in a TED Talk, "Americans like instant gratification.") The book was one of the first explorations of what are now called existential risks — threats, especially technological ones, that could plausibly lead to humankind's extinction. Rees argued that we are living through a critical moment in the history of our species, and that we will either find a way to responsibly harness these technologies or be destroyed by them.

In *On the Future*, Rees casts a wider eye over humankind, which he sees threatened by powerful new technologies like gene editing and the accumulating pressures of growth and climate change. In a Skype conversation from his home in Cambridge, England, Rees discussed the limits of science, his political pessimism, and why humans of the future could be very different from you or me.

This interview has been lightly edited and condensed for clarity.

Medium: In "*Our Final Century,*" you focused on the existential threats facing our species, especially from emerging technologies. You touch on that subject again in "*On the Future*." How has your level of concern changed over the years?

Martin Rees: I don't think it's changed, really. I describe myself as a technological optimist but a political pessimist. We have the technology to provide a good life for 9 billion people in the world by the middle of the century, but we'll have a bumpy ride because of political constraints. That's still my view.

The concerns I raise in my new book are of two kinds. One focuses on the pressures we're putting collectively on the biosphere, because there are more of us, all demanding more energy and resources, which leads to climate change, loss of biodiversity, and so forth. That's one class.

> Even an individual can, by error or by design, cause a catastrophe that cascades very widely, even globally.

But the other concern focuses on new technologies, like in bio, cyber, and A.I., which are, of course, very powerful. They've developed hugely over the last 15 years. And they lead to a new issue of concern that just a few people or even an individual can, by error or by design, cause a catastrophe that cascades very widely, even globally. This is something which is very new. And this is a big challenge, because it adds to the tension between privacy, security, and liberty, if you want to minimize that risk. I like to say the global village will have its village idiots, and they will have a global range. So that's a new concern I have.

If we can manage these new technologies and the risks around growth and climate change, will we move to a safer period? Or will these existential risks always be with us?

I don't think so. My main theme is that if we think in the cosmic perspective—and I'm an astronomer by profession—we know that the Earth has existed for 45 million centuries. And this one species, the human species, can determine the future of the planet. Of course, there are doomsday scenarios where we leave a denuded planet and have mass extinctions, and there are bright scenarios where we survive on this planet and perhaps spread beyond it. We can initiate a transition from humans to some post-human species, augmented genetically or augmented by cyborg techniques.

You describe yourself as a technological optimist but a political pessimist. Is that chiefly a function of our times, or has that always been your outlook?

Well, I think no one could look at the global situation now without being a political pessimist. But the way I see it is if we look back at the Middle Ages—say, 500 years ago—things in Europe were far, far worse than they are now. We can't deny there has been a huge amount of technical development, without which we would not be able to have a world of 7 billion people where we avoid mass starvation. Five hundred years ago, things were miserable, but there wasn't much that could be done about it.

On the other hand, the gap today between the way things are and the way they could be is wider. Most of us are doing fine, but we also know that the richest 1,000 people in the world could make a big difference to the lives of the bottom 1 billion in the world. The fact that they're not doing that is, of course, an ethical indictment of this present situation. That's why I question the idea of real ethical progress over that time.

On a threat like climate change, there are those who will

put forward primarily political solutions — carbon taxes and caps — and those who will focus on technological solutions. Which route will we find success with: the political or the technological?

I think the latter is the only one with any chance of success, because it's a very hard sell, politically, to persuade people to make sacrifices now, say, to take on a high carbon tax, when the benefits mainly accrue several decades in the future and mainly to other parts of the world. Politicians focus on the local and the short term, and it's very hard to get them to adopt long-term issues.

On the other hand, I think we could do far more to accelerate research and development into all kinds of clean energy, smart continent-wide grids, and energy storage. And this is a win-win situation — a win for the high-tech nations which develop it, and more importantly, a win for countries like India that now get energy from smoky stoves burning wood and dung. They need to have a grid. They need more power. What they're tempted to do is to adopt the cheapest energy, which is coal-fired power stations. But we want to accelerate R&D so that gradually the costs come down and they can afford to leapfrog directly to clean energy.

Are there potential limits to scientific progress, questions that we might never be able to answer?

There are certainly often questions which aren't really tractable. To give one example from my own area of interest, there's the origin of life. We know about evolution, but the actual origin of life, the transition from chemistry to the first metabolizing reproducing system that we call life, that's something which everyone has known for at least 50 years is a crucial problem, but no one has really known how to tackle it.

I think your question has another aspect — are there some parts of science which we'll never be able to tackle because our brains aren't up to it? I think we need to be open-minded about that, too. There's no particular reason to think that our brains, which haven't changed much since our ancestors roamed the African savanna, can cope with understanding all of physical reality. It's surprising enough that we can cope with the counterintuitive world of the quantum and of the cosmos, away from everyday scales and everyday physics.

But there may be some things that we'll never be able to understand. Computers might help us, but we'll never have full insight into some sciences, I suspect. This could be true of some fundamental physics like string theory, which may be correct but too difficult for us to actually test and verify. And it may also be true of trying to fully understand our own brains.

Whether it's biotechnology and the possibility of viruses that could be engineered to extreme lethality or the rise of artificial intelligence or nanotechnology, what existential risks keep you up at night?

If you think of the immediate situation, cyber threats are very concerning. We know they're happening already. We know that there's an arms race between those who are trying to carry

out cyberattacks and those who are trying to protect against them. We know from the U.S. Department of Defense that a single person could carry out a cyberattack, for example, on the electrical grid that is so severe as to demand a nuclear response. They actually used those words. So cyberthreats are very, very serious.

In the short term, I also worry about bio threats, by error or by terror. It's possible for a few people to create some sort of a dangerous pathogen that could be released by error or by design. There have been attempts to get possible regulations. That's very good and very important, but sometimes I worry that those regulations can't now be enforced because these technologies have been pursued globally with strong commercial pressures. And I worry that enforcing them globally is as hopeless as enforcing drug laws globally or tax laws globally. We've had precious little success in either of those. I worry that anything that can be done will be done somewhere by someone.

What about A.I. in particular?

As regards A.I. and the machines taking over, of course some people worry about that. But I think that is a longer-term concern, and I think it may be misleading. People assume that machines will be aggressive and expansionist like we are. We have evolved via natural selection, by a process which has favored intelligence but also favored aggression. But if Darwin and biological evolution is replaced by technological evolution, there's no particular reason why it should be strictly analogous to the harsh survival of the fittest. I don't see why we would assume that these machines will be aggressive rather than contemplative.

Is there anything we should be doing now to prepare for the possibility of artificial superintelligence?

There are those who think that for a long time we need to worry far more about human stupidity than artificial intelligence and that it's premature to have these discussions now. You have to bear in mind that although it's hugely impressive what A.I. has done in special areas like playing games, robots are still rather primitive compared to humans when it comes to interacting with the external world.

But the order in which things happen with A.I. will make a big difference. For instance, you can develop an A.I. which is very powerful, and it can be hugely helpful, not only in controlling complex networks of traffic flow, electricity grids, and things like that, but also perhaps in helping us address some of these other threats we've discussed. So we want to develop the right kind of A.I. very quickly to help us.

> People assume that machines will be aggressive and expansionist, like we are.

On the other hand, we want to ensure that we don't accidentally develop A.I. which could go rogue and interact via the internet of things with the external world.

In recent years, we've seen the establishment of organizations focused on existential risk, like the Future of Humanity Institute at Oxford or the Centre for the Study of Existential Risk at Cambridge, which you are

affiliated with. Have these organizations been effective when it comes to raising awareness of existential risks or even formulating strategies for how we should address them?

I think they have been useful. A huge amount of effort goes into reducing more conventional risks, things like low radiation doses, aircraft accidents, and so forth. In comparison with those kinds of risks, these new ultrahigh-consequence but low-probability threats are very under-discussed. Even with these institutes, worldwide there are probably only about 100 or so people who would claim that their prime interest is in addressing existential threats.

Given how huge the stakes are, if these people can reduce the probability of an existential catastrophe by just one part in 10,000, they've more than earned their keep. I think that there should be rather more people addressing these long-term threats.

Do you think that in 200 years time, humanity will be in a much better place than it is now? Will we still even be here?

I'm worried, because I think we can't predict. We have a bumpy ride ahead of us, because things will go wrong. There are going to be major political errors. But I would hope that in the long run, we will survive.

As an astronomer, do you believe it is human destiny for us to end up in space?

I have some slight heretical views about manned space flight. I think because robots are getting better, the practical case to send people into space is getting weaker all the time. If I was an American, I wouldn't support the manned part of NASA's program at all. I'd support the unmanned part, but the manned part I would leave to Elon Musk and Jeff Bezos.

They can take higher risks than NASA could impose on civilian astronauts. They can fly the kind of adventures with astronauts who are prepared to take a 10 or 20 percent risk of not coming back. I think that's what will happen, and I think we should cheer them on. Through their efforts, there will be some people who, by the end of the century, will be living on Mars. But I don't think it will be mass emigration. It's a dangerous delusion to believe we can solve the problems of Earth by going to Mars.

Dealing with climate change is a doodle compared to terraforming Mars. Those who go will have every incentive to adapt themselves to a hostile environment using genetic modification and cyborg techniques which may be highly regulated here on Earth.

So the future of humanity may mean going beyond humanity?

It could be that, within a few hundred years, they will have evolved into a new species, perhaps one that is electronic and near immortal. Post-human evolution may be

spearheaded by the few thrill-seeking pioneers who go to Mars. They can then go off into the blue yonder and pervade interstellar space.