The New York Times
Opinionator

JANUARY 27, 2013, 5:00 PM

Cambridge, Cabs and Copenhagen: My Route to Existential Risk

By ***HUW PRICE***

In Copenhagen the summer before last, I shared a taxi with a man who thought his chance of dying in an artificial intelligence-related accident was as high as that of heart disease or cancer. No surprise if he'd been the driver, perhaps (never tell a taxi driver that you're a philosopher!), but this was a man who has spent his career with computers.

Indeed, he's so talented in that field that he is one of the team who made this century so, well, 21st - who got us talking to one another on video screens, the way we knew we'd be doing in the 21st century, back when I was a boy, half a century ago. For this was Jaan Tallinn, one of the team who gave us Skype. (Since then, taking him to dinner in Trinity College here in Cambridge, I've had colleagues queuing up to shake his hand, thanking him for keeping them in touch with distant grandchildren.)

I knew of the suggestion that A.I. might be dangerous, of course. I had heard of the "singularity," or "intelligence explosion"- roughly, the idea, originally due to the statistician I J Good (a Cambridge-trained former colleague of Alan Turing's), that once machine intelligence reaches a certain point, it could take over its own process of improvement, perhaps exponentially, so that we humans would soon be left far behind. But I'd never met anyone who regarded it as such a pressing cause for concern - let alone anyone with their feet so firmly on the ground in the software business.

I was intrigued, and also impressed, by Tallinn's commitment to doing something about it. The topic came up because I'd asked what he worked on these days. The answer, in part, is that he spends a lot of his time trying to improve the odds, in one way or another (talking to philosophers in Danish taxis, for example).

I was heading for Cambridge at the time, to take up my new job as Bertrand Russell professor of philosophy - a chair named after a man who spent the last years of his life trying to protect humanity from another kind of technological risk, that of nuclear war. And one of the people I already knew in Cambridge was the distinguished cosmologist Martin Rees - then master of Trinity College, and former president of the Royal Society. Lord Rees is another outspoken proponent of the view that we humans should pay more attention to the ways in which our own technology might threaten our survival. (Biotechnology gets most attention, in his work.)

So it occurred to me that there might be a useful, interesting and appropriate role for me, as a kind of catalyst between these two activists, and their respective circles. And that, to fast forward a little, is how I came to be taking Jaan Tallinn to dinner in Trinity College; and how he, Martin Rees and I now come to be working together, to establish here in Cambridge the Centre for the Study of Existential Risk (C.S.E.R.).

By "existential risks" (E.R.) we mean, roughly, catastrophic risks to our species that are "our fault," in the sense that they arise from human technologies. These are not the only catastrophic risks we humans face, of course: asteroid impacts and extreme volcanic events could wipe us out, for example. But in comparison with possible technological risks, these natural risks are comparatively well studied and, arguably, comparatively minor (the major source of uncertainty being on the technological side). So the greatest need, in our view, is to pay a lot more attention to these technological risks. That's why we chose to make them the explicit focus of our center.

I have now met many fascinating scholars - scientists, philosophers and others - who think that these issues are profoundly important, and seriously understudied. Strikingly, though, they differ about where they think the most pressing risks lie. A Cambridge zoologist I met recently is most worried about deadly designer bacteria, produced - whether by error or by terror, as Rees puts it - in a nearby future in which there's almost an app for such things. To him, A.I. risk seemed comparatively far-fetched - though he confessed that he was no expert (and added that the evidence is that even experts do little better than chance, in many areas).

Where do I stand on the A.I. case, the one that got me into this business? I don't claim any great expertise on the matter (perhaps wisely, in the light of the evidence just mentioned). For what it's worth, however, my view goes like this. On the one hand, I haven't yet seen a strong case for being quite as pessimistic as Jaan Tallinn was in the taxi that day. (To be fair, he himself says that he's not always that pessimistic.) On the other hand, I do think that there are strong reasons to think that we humans are nearing one of the most significant moments in our entire history: the point at which intelligence escapes the constraints of biology. And I see no compelling grounds for confidence that if that does happen, we will survive the transition in reasonable shape. Without such grounds, I think we have cause for concern.

My case for these conclusions relies on three main observations. The first is that our own intelligence is an evolved biological solution to a kind of optimization problem, operating under very tight constraints of time, energy, raw materials, historical starting point and no doubt many other factors. The hardware needs to fit through a mammalian birth canal, to be reasonably protected for a mobile life in a hazardous environment, to consume something like 1,000 calories per day and so on - not to mention being achievable by mutation and selection over a time scale of some tens of millions of years, starting from what existed back then!

Second, this biological endowment, such as it is, has been essentially constant, for many thousands of years. It is a kind of fixed point in the landscape, a mountain peak on which we have all lived for hundreds of generations. Think of it as Mount Fuji, for example. We are creatures of this volcano. The fact that it towers above the surrounding landscape enables us to dominate our environment and accounts for our extraordinary success, compared with most other species on the planet. (Some species benefit from our success, of course: cockroaches and rats, perhaps, and the many distinctive bacteria that inhabit our guts.) And the distinctive shape of the peak - also constant, or nearly so, for all these generations - is very deeply entangled with our sense of what it is to be us. We are not just creatures of any volcano; we are creatures of this one.

Both the height and the shape of the mountain are products of our biological history, in the main. (The qualification is needed because cultural inheritance may well play a role too.) Our great success in the biological landscape, in turn, is mainly because of the fact that the distinctive intelligence that the height and shape represent has enabled us to control and modify the surrounding environment. We've been exercising such control for a very long time of course, but we've recently got much better at it. Modern science and technology give us new and extraordinarily powerful ways to modify the natural world, and the creatures of the ancient volcano are more dominant than ever before.

This is all old news, of course, as is the observation that this success may ultimately be our undoing. (Remember Malthus.) But the new concern, linked to speculation about the future of A.I., is that we may soon be in a position to do something entirely new: to unleash a kind of artificial vulcanism, that may change the shape and height of our own mountain, or build new ones, perhaps even higher, and perhaps of shapes we cannot presently imagine. In other words - and this is my third observation - we face the prospect that designed nonbiological technologies, operating under entirely different constraints in many respects, may soon do the kinds of things that our brain does, but very much faster, and very much better, in whatever dimensions of improvement may turn out to be available.

The claim that we face this prospect may seem contestable. Is it really plausible that technology will reach this stage (ever, let alone soon)? I'll come back to this. For the moment, the point I want to make is simply that if we do suppose that we are going to reach such a stage - a point at which technology reshapes our human Mount Fuji, or builds other peaks elsewhere - then it's not going to be business as usual, as far as we are concerned. Technology will have modified the one thing, more than anything else, that has made it "business as usual" so long as we have been human.

Indeed, it's not really clear who "we" would be, in those circumstances. Would we be humans surviving (or not) in an environment in which superior machine intelligences had taken the reins, to speak? Would we be human intelligences somehow extended by nonbiological means? Would we be in some sense entirely posthuman (though thinking of ourselves perhaps as descendants of humans)? I don't claim that these are the only options, or even that these options are particularly well formulated - they're not! My point is simply that if technology does get to this stage, the most important fixed point in our landscape is no longer fixed - on the contrary, it might be moving, rapidly, in directions we creatures of the volcano are not well equipped to understand, let alone predict. That seems to me a cause for concern.

These are my reasons for thinking that at some point over the horizon, there's a major tipping point awaiting us, when intelligence escapes its biological constraints; and that it is far from clear that that's good news, from our point of view. To sum it up briefly, the argument rests on three propositions: (i) the level and general shape of human intelligence is highly contingent, a product of biological constraints and accidents; (ii) despite its contingency in the big scheme of things, it is essential to us - it is who we are, more or less, and it accounts for our success; (iii) technology is likely to give us the means to bypass the biological constraints, either altering our own minds or constructing machines with comparable capabilities, and thereby reforming the landscape.

But how far away might this tipping point be, and will it ever happen at all? This brings me back to the most contested claim of these three - the assertion that nonbiological machines are likely, at some point, to be as intelligent or more intelligent than the "biological machines" we have in our skulls.

Objections to this claim come from several directions. Some contest it based on the (claimed) poor record of A.I. so far; others on the basis of some claimed fundamental difference between human minds and computers; yet others, perhaps, on the grounds that the claim is simply unclear - it isn't clear what intelligence is, for example.

To arguments of the last kind, I'm inclined to give a pragmatist's answer: Don't think about what intelligence is, think about what it does. Putting it rather crudely, the distinctive thing about our peak in the present biological landscape is that we tend to be much better at controlling our environment than any other species. In these terms, the question is then whether machines might at some point do an even better job (perhaps a vastly better job). If so, then all the above concerns seem to be back on the table, even though we haven't mentioned the word "intelligence," let alone tried to say what it means. (You might try to resurrect the objection by focusing on the word "control," but here I think you'd be on thin ice: it's clear that machines already control things, in some sense - they drive cars, for example.)

Much the same point can be made against attempts to take comfort in the idea that there is something fundamentally different between human minds and computers. Suppose there is, and that that means that computers will never do some of the things that we do - write philosophy, appreciate the sublime, or whatever. What's the case for thinking that without these gifts, the machines cannot control the terrestrial environment a lot more effectively than we do?

People who worry about these things often say that the main threat may come from accidents involving "dumb optimizers" - machines with rather simple goals (producing IKEA furniture, say) that figure out that they can improve their output astronomically by taking control of various resources on which we depend for our survival. Nobody expects an automated furniture factory to do philosophy. Does that make it less dangerous? (Would you bet your grandchildren's lives on the matter?)

But there's a more direct answer, too, to this attempt to take comfort in any supposed difference between human minds and computers. It also cuts against attempts to take refuge in the failure of A.I. to live up to some of its own hype. It's an answer in two parts. The first part - let me call it, a little aggressively, the blow to the head - points out that however biology got us onto this exalted peak in the landscape, the tricks are all there for our inspection: most of it is done with the glop inside our skulls. Understand that, and you understand how to do it artificially, at least in principle. Sure, it could turn out that there's then no way to improve things - that biology, despite all the constraints, really has hit some sort of fundamental maximum. Or it could turn out that the task of figuring out how biology did it is just beyond us, at least for the foreseeable future (even the remotely foreseeable future). But again, are you going to bet your grandchildren on that possibility?

The second part of the argument - the blow from below - asks these opponents just how far up the intelligence mountain they think that A.I. could get us. To the level of our fishy ancestors? Our early mammalian ancestors? (Keep in mind that the important question is the pragmatic one: Could a machine do what these creatures do?) Wherever they claim to draw the line, the objection challenges them to say what biology does next, that no nonbiological machine could possibly do. Perhaps someone has a plausible answer to this question, but for my part, I have no idea what it could be.

At present, then, I see no good reason to believe that intelligence is never going to escape from the head, or that it won't do so in time scales we could reasonably care about. Hence it seems to me eminently sensible to think about what happens if and when it does so, and whether there's something we can do to favor good outcomes over bad, in that case. That's how I see what Rees, Tallinn and I want to do in Cambridge (about this kind of technological risk, as about others): we're trying to assemble an organization that will use the combined intellectual power of a lot of gifted people to shift some probability from the bad side to the good.

Tallin compares this to wearing a seat belt. Most of us agree that that makes sense, even if the risk of an accident is low, and even though we can't be certain that it would be beneficial, if we were to have an accident. (Occasionally, seat belts make things worse.) The analogy is apt in another way, too. It is easy to turn a blind eye to the case for wearing a seat belt. Many of us don't wear them in taxis, for example. Something - perhaps optimism, a sense that caution isn't cool, or (if you're sufficiently English!) a misplaced concern about hurting the driver's feelings - just gets in the way of the simple choice to put the thing on. Usually it makes no difference, of course, but sometimes people get needlessly hurt.

Worrying about catastrophic risk may have similar image problems. We tend to be optimists, and it might be easier, and perhaps in some sense cooler, not to bother. So I finish with two recommendations. First, keep in mind that in this case our fate is in the hands, if that's the word, of what might charitably be called a very large and poorly organized committee - collectively shortsighted, if not actually reckless, but responsible for guiding our fast-moving vehicle through some hazardous and yet completely unfamiliar terrain. Second, remember that all the children - all of them - are in the back. We thrill-seeking grandparents may have little to lose, but shouldn't we be encouraging the kids to buckle up?

Huw Price is Bertrand Russell professor of philosophy at the University of Cambridge. With Martin Rees and Jaan Tallinn, he is a co-founder of the project to establish the Centre for the Study of Existential Risk.