AI and You

Transcript

Guest: Tony Czarnecki, part 1

Episode 133

First Aired: Monday, January 2, 2023

Hello, and welcome to episode 133! Returning to the show today is Tony Czarnecki. Tony was first on the show in episodes 18 and 19 in October 2020, and he's had plenty of time in which to come up with new and fascinating thoughts to discuss with us. Tony is a futurist in the United Kingdom, a member of the Chatham House in London, and the Managing Partner of Sustensis, also in London – a think tank for inspirations for humanity's transition to coexistence with superintelligence. Tony is the author of several books on the subject of superintelligence, three of which form the *Posthumans* series. His latest paper is called, "How might Transhumans control Superintelligence?" and there is a provocative title if ever I heard one. Of course that's all right in our wheelhouse for this podcast.

If the term *transhuman* is new or relatively new to you, here's some explanation. Transhumanism is concerned with the next stage of human evolution or development – in other words, the transformation of the human species into something more advanced. That's quite general, of course, and in practice it tends to get narrowed down into various ideas and factions focused on particular kinds of transformation, whether they be nanobots or genetic engineering or the use of electromechanical implants to augment human capability, and of course these groups often quarrel with each other or attract some negative press, because, after all, most people hear a description like that and start thinking about the Borg on Star Trek or some other dystopian cautionary tale. The idea that we shouldn't meddle with the hand fate has dealt us runs deep.

In our conversation the term transhuman is more of a nod to the inevitability that the arrival of superintelligence would mean some kind of transformation of the human race than what specific form that transformation would take, other than it would necessarily involve the ability to communicate rapidly with superintelligent AI through some kind of brain-computer interface. And I think this is a compelling idea worthy of a lot of our attention, that the more AI evolves, the closer we get to needing to evolve ourselves, in ways that we really need to consider a lot more closely. Some people do consider it; and one of those is Tony Czarnecki, so let's get right to the interview.

Tony, welcome to the show.

Thank you. Nice to meet you again, Peter.

Welcome back, I should say; and last time, we talked at length about mostly international socioeconomics and how it would be affected by AI and now you've got a new paper, published in August 2022, quite recently, called: "How Might Transhumans Control Superintelligence?" with the subtitle "Will 2030 be a Tipping Point for Artificial Intelligence Control" and so the words that I really wanted to focus on there are "transhuman," "control," "superintelligence," and "2030," in other words, all of it, because they really deserve some definition and amplification and actually, one of the things that strikes me first as 2030 is not that far away

now. We're eight years away from that and that puts it within a certain horizon, once we get to predicting timeframes more than 10 years out, people tend to get very, when it comes to technology; fluid in their predictions, they think, well, a lot of things are possible in 10 years, and in particular, in 20 years, we think that just about anything is possible. But here you are within an eight-year timeframe, which is, puts it in the sort of frame where thing people like fund managers start thinking, oh, well, maybe I should change my investment profiles. So, first of all, I'm going to look at that timeframe. But we should look, first of all, what we're talking about happening in that timeframe. So, let's talk about your tipping point that you see. In that timeframe. What do you mean by "tipping point"?

Yes, I think it's right to start with 2030 because it's a purposely a provocative title to make people thinking, yes, it's just eight years away, how on earth this man is suggesting that we may lose control of our AI that we just develop, or essentially, over our destiny. So, let me just give you some parse. You may remember what kind of fight the Green lobby was fighting in until 2015, almost for 30 years to convince the general public that global warming is real and not only that, that if we don't do anything substantial to reduce the global warming below 1.5 centigrade by 2030 precisely, then we may be proverbially cooked, almost literally. So, this is my time horizon as well, but not just for global warming, but to other, I would say global threats, existential threats. The second one is what I call global disorder and the mark 2030 was penciled by the United Nations Millennium Project, which is to deliver by 2030 17 so called sustainability goals. It is addressed primarily towards the poorer countries, the developing countries, and the understanding behind the project is that if nothing significant happens in those countries that means the improvement of the social and economic standards, those people who mainly live in the southern hemisphere, where the drought may be especially prominent, may have nothing to lose, and they will just march towards the northern hemisphere and that may start, together with other existential risks, such as pandemic and local nuclear wars, if you like or even local, non-nuclear wars, it may create a global disorder that altogether will create an existential risk for humans. So, this is the second 2030 date. The third date is for artificial intelligence. So far, there are only very few people that ought to stick their head, so to speak above the parapet, who say that by 2030, we may lose that control, either directly or indirectly, among others. One of the most prominent futurists, Ray Kurzweil, says that it may happen by 2029. Just recently, he confirmed that date, although he's not so sure whether it's not going to be earlier than that. So, I'm always on that cautionary side saying that 2030 may be the date.

And I just want to interrupt there, because I think Kurzweil is not saying that we will lose control of AI by 2030.

No.

He's saying we will have human level AI by 2029, which he is quite optimistic about. I just wanted to interject that.

Yes, that's why I'm saying, "indirectly." But I will go even further, saying that, if such a date is not established, nothing substantial regarding the AI control will happen, as it didn't happen with global warming.Yes, I have no scientific proof that by 2030, we may lose control over AI, as the

IPCC doesn't have any scientific proof that by 2030, we will have a global average temperature increase over 1.5 centigrade. This is only a best guess and so mine is also the best guess. But it is, in my view, pretty probable, and better be safe than sorry and prepare for the worst. So that will be my justification why I'm putting this 2030 as the tipping point, to concentrate the efforts so that we will advance any necessary controls of AI on time.

Well, let's take our time and look at just what that AI might be at that point in 2030, needs this level of attention given to it. So, first of all, are you in agreement with Kurzweil's estimate? He says 2029 for human level artificial Intelligence?

Yes, definitely, I am in line with his thinking. I would also be prepared that this might happen earlier. I mean, this year has been absolutely spectacular regarding discoveries of AI. None of the years before has had so many profound new discoveries, new capabilities of AI, including the first steps toward cognition, using videos and so on. We are on a very dangerous path already. Imagine that we have somehow - or AI would have helped us - deliver some sort of cognitive AI by 2030. Then all the calculations go out the window, because a cognitive AI would come to its own "conclusions" what it can do. It will try almost to play with the constraint that it will have it to see whether it can bypass. It will by then have the universe of internet connected items, IOTs on a much larger scale than now. Just think about millions of Optimus robots, that Mr. Musk has just confirmed he's going to deliver like Tesla cars in millions costing around $20,000, even if they're not perfect. But if they have access to the Internet, right, they can interconnect then with each other in any part of the world. And this is only one of, I'd say, dozens and dozens of examples that can create such a network of unintended skilfull networks which may which may run away on their own.

In your paper, you refer to the superintelligence of that year 2030. You refer to that as an "immature" superintelligence. What makes it immature?

Correct, Yes, that's a very good, unnecessary quarter. vacation, I am not saying that we will have super intelligence by 2030. It's rather 2045, I would say 2050 as Kurzweil has suggested. Well, I think we have right now is an Artificial Narrow Intelligence, that means an intelligence that is very capable, in many cases thousands of times more capable than any human, and then entirely stupid in all the other areas. And that dichotomy will only grow. And because of that, I call it an immature AI, like a teenager, that will try to manipulate the nuclear bomb, just for fun. Artificial intelligence in eight years' time will be far, far more so, perhaps thirty times more intelligent than it is now. So just think about that. I'll leave it there because I could spend more time to justify my supposition.

Well, there are several different terms that sound like they might be the same ones, that I want to see whether they really are. So Kurzweil talks about human level artificial intelligence, which you've also been talking about, and superintelligence. And then another common term is artificial general intelligence; and is artificial general intelligence the same as human level artificial intelligence?

Yes, and no and thank you for asking to define those terms, because people are completely lost in terms of terminology. Not even those who are in AI agree on those terms, only now I can see for the last six months or so, people are gradually going into the same directions calling the same things the same. For me, Artificial Narrow Intelligence is that's what we have right now. Which is working on the probability of reaching its goals and not uncertainty like IT. It can operate without all the input data being available, which differs it from IT as well. But this artificial intelligence doesn't have cognition, it doesn't know what it says. Even if it writes incredibly accurate scientific reports like GPT-3, prompted by questions, it doesn't know what it's writing, it doesn't have a clue of any sentence that it says. However, we have to be more humble and we have to understand that the intelligence as we define, as humans define, doesn't have to always be the human in the way; actually intelligence may be derived in a different way and what GPT-3 or LaMDA, perhaps the most notable and natural language processors, what they tell us that the product of the encodes thinking, it is equal to best scientists to such a degree that there are now companies that sell such reports generated by GPT-3 or LaMDA and you can argue whether they have the right to do so. So, the intelligence that in charge, does deliver value if you like. Now, regarding whether it is possible to control single or Narrow Intelligence, it depends on the, I think, on the spread of that intelligence and how well it is globally interconnected. At the moment, we are I think reasonably safe, but I think even in the few years' time, where Starlink is in full operation, and when we will have millions of intelligent robots, humanoid robots, this situation will be blurred and we may not know where the answers have come, from which network or from which AI network, how it has generated ideas or how it has generated decisions. This is the black box scenario. Thankfully we now have probably a tool to discern the way that AI gets its answers. It's so good the white box delivered in France in May this year. Very rarely people are talking about, but I think it's a very important step in direction of controlling AI. But in summary, this AI, today's AI, artificial, Narrow Intelligence, by the end of this decade, may indeed be dangerous. I'm not saying that will be completely out of control. It's probably unlikely. The date 2030 is the most hypothetical, but it may. I think the probability is still in the lower figures. But it may be just a few years later. So, I'm not going to struggle to reassess and rejustify why 2030; I would say about; most likely they'd like the global warming tipping point.

Right. When it comes to something as big a threat as what you're talking about, we shouldn't quibble over a couple of years either way, we should start getting ready. And I just want to interject when you're talking about the explainability problem there on white boxes that I want to refer listeners to an earlier episode with Michael Hind of IBM, who was an expert on explainability and talked about several really intriguing ways and interesting research into how artificial intelligence can explain its decisions today. I want to go back to when you were talking about an AI playing with nuclear bombs just because it would be interesting. And that's a level of anthropomorphization that's familiar through Hollywood tropes, in particular science fiction tropes, that communicates to us, it powerfully suggests, something that is conscious, human-like in the way that it frames goals, and has survival instincts and other things about it and all of that cascades to us and causes certain thinking about how that's going to happen. Do you see in that scenario that such an AI is necessarily conscious?

If we are talking about the ultimate, artificial general intelligence, and that would be maturing towards superintelligence with the Singularity being the ultimate escape route for it, then I think it will be conscious.

It is the one that you talked about playing with the bomb than that, by that point, is that necessarily conscious?

Doesn't have to be, in my view.

And what does play with a bomb, what form does that take? Then that sort of activity without being conscious? Yes. How does that happen?

Let's come back to your initial question for a moment when you asked me and I didn't answer it fully, what I mean by artificial general intelligence, as Ray Kurzweil implies, human level intelligence by 2029. I visualize it, not so much as a global system; it may be partially global, I'm rather thinking about a humanlike almost, that will have skills and there has to operate in a real environment, it cannot sit within a computer like that famous sci-fi film *Her*, right? It is an avatar with its own brains, and own legs and so on that interacts with an environment and may interact with other maybe 1000s of even millions of its brethren. So such an artificial general intelligence is limited by not being conscious; may be cognitive, but I think it's a little bit iffy. I think it is likely that it will be cognitive. So, it will know its environment, and it will know that it can't go through the wall, right? It will know enough about the environment as a kid knows, say at that level but it will not know the relationship, not to mention the emotions that humans go through. But the superintelligence, the ultimate version of artificial intelligence, I think will have emotions and will be conscious. And the reason why I'm saying this, let's talk about emotions. First, that emotions are nothing else [than] an electromagnetic wave process. It's nothing else, we can induce it, even now in ourselves, and we can record it and so. So, similarly, I'm in the same camp as John McFadden, who is the father of electromagnetic-field-induced consciousness. I'm not going to go into the details, because it may bore somebody but it is possible to visualize that consciousness may be present in non-biological entities. So, AGI in 2030. This is, let's say, Optimus super intelligent in this sense that it can play a violin, it can cook a dinner and some will read us a book, which is easy even now, and so on. But it will not be a global system with the satellite connection so that it would overpower it, it have still a lot to learn, especially in the emotional area, relationship area, that may be very difficult.

Would it be fair to say then that at this early stage - the 2030 or so stage - that AI could be convincingly and usefully faking kind of general intelligence, in the same way that right now we have large language models that are doing a convincing job of faking being an artist. You ask it to draw something novel that no one has ever drawn or seen before, and it does really well straight away and so that's producing the same results as something that was creative, even if we want to argue that it isn't, and even if we want to argue that it's not artistic, and we sort of crossed the line with Blake Lemoine, claiming that LaMDA was sentient, no one agrees with him. But we can see that we could cross that line more and more with other people saying, it may not be sentient on the inside, but it's doing a good enough job on the outside. So, could

that mean then that you're eventually, we get to around 2030 and AI is doing a good job of faking actions like playing around, as you put it, on this sort of global scale?

Yes. Spot on. Absolutely. The Blake Lemoine incident is very teaching and a little bit scary, to some extent, but I don't believe in any second that it had any resemblance of consciousness. It couldn't because it was sitting within two-dimensional screen. So, that's for now. But around 2030, IIthink we quite often wouldn't know whether the artificial intelligent agents and humanoid does really feel something emotional, or whether it pretends. We already have such pretending robots today. Like Ameca. So, yes, we can hardly distinguish what is real, or when watching TV, when you have holograms in augmented reality, like "Strictly Come Dancing," which I don't watch anymore, but you still and you are not sure what is real, what is fake, or what is a hologram.

Tell us about the Ameca robot. I don't know that one.

Yeah. Well, what is fascinating about that robot was created in around six months by a British company from Bristol that used to produce robots for Disneyland and so on for 20 years. And suddenly, they discovered that they could do it on a big AI stage, and they did it and just this month, they released Ameca 2 with GPT-3 embedded. I saw only some clips, there are not too many yet, how it reacts. So, we are moving in that direction. Natural language processing will be bundled with the hardware. So this knowledge that only some of us have access to will be suddenly available at the street level. That will be an experience within next 12 months, maybe two years at most. Think about the Christmas time, right? Those guys inviting people in the Harrods' store.

Right. And I know though the one you mean: this is the robot that a few months ago I saw the videos where it has a quite expressive face, it's able to do very convincing facial emotions and GPT-3 is the engine behind that. So; tipping point: is the tipping point the hard takeoff, is the tipping point recursive self-improvement where the AI gets to the point where it can teach itself out how to learn faster? Or was it an earlier one?

I made a number of points. Or if you like, milestones, by which we may be judging how far AI is escaping out of our control. And there are a number of them. One may be related to hardware, for instance, that an artificial humanoid can have as many neurons as a human – eighty-six billion and by the current progress, it may happen in two years' time. 3000 artificial neurons are roughly equivalent and equivalent of one human neuron. So, you need trillions of neurons to have 86 billion of human equivalent. So, that will be one measure. Because then the capacity, that's what Kurzweil is talking about the capacity of its brain will equal our brain. Doesn't mean that it's the same or it can do the same things. Because for that, it we have to consider that each neuron is connected to roughly almost 10,000 other neurons, which may take years to develop it. But at some stage, as you said, Peter, that we may come to a situation where we will not know whether what we are being told is everything that that AI knows. And when it is proven, then that has happened, that will be next milestone, right? It may not still be out of our control, but it will be on the runaway pathway. The third one may be that it will create a global chaos. I would put it at around 2027, 28, in five-years' time. By error, almost obviously, not intentionally, yet.

Which will create, I don't know, a panic on the stock exchanges and even in military sense, and so on. That I think will happen in just five years' time, and I am so bold about it because I can see how it can be made, judging the progress that has been made in the last few years, even this year, which is so scary that we've made such an enormous progress in just one year in many disciplines, on the physics and so on.

So, for the significant fraction of our audience right now who are approaching panic and getting visions of Optimus robots running amok, like the movie *I, Robot*. What sort of crisis do you foresee in that timeframe? How do you imagine that happening?

The example that I just given, I think they may be repeated a few times by the end of this decade, and its severity is proportional to our preparedness to face such challenges. And you know, Peter, why I am really upset that so little is done by that, because most politicians and decision-makers believe that change is happening at the same pace as it used [to] for the entire duration of the humankind. Whereas for the last few years, in many domains, such as medicine, the change is progressing at almost exponential pace and because of that, when such an incident happens, the government and the emergency services working at, if you like, in the same mode, assuming linear progression of change, they will come too late [to] do anything that is worthwhile doing because the damage will be done. And this is so real and it can be proven so convincingly in laboratories and so on if you like within even today, or maybe two three years' time, and we should be prepared, we should be very open about that. I don't want to scare it. I'm an optimist, generally as you probably know, I'm a real optimist, I have to look for what may actually stop it happening. How can we remain in control?

And I want to point out another reason that people in charge, to coin a phrase, don't act in the way that you'd like, or that we would like, and it was pointed out by Oliver Letwin, a British politician, in his address to a Center for the Study of Existential Risk conference last year, I believe, when he said, for even for those of us who realize the risks of everything from pandemic to AI revolt, if we try taking prophylactic action on that now, we would just be kicked out of office; it is not something where we can convincingly sell that to an electorate who will wondering why aren't you spending that time lowering my energy bills?

I can't agree more. This is precisely the point I'm trying to make, how ill-prepared we are. But on the positive side, I am in touch with a UN-related organization. I won't mention right now, which is preparing a conference on controlling AI with the vision to create an organization in 10 years' time, right? And that is, if you like, an ordinary way of how things are being done. And one may say, especially the politicians, look it's time, even if it is so dangerous, we can't do it over the next year and so on we need time. For to which I respond to like this: Do you know how long it took to create the United Nations? Two years. Do you know how long it took to create NATO? One year. Now you may ask another question. Why took NATO to be such a - powerful organization anyway - and meaningful organization? How it took just one year to create? The answer is because this was almost just after the war. People remembered what it was and it was a global threat, a threat of nuclear war. What is dissimilar about the current situation? Nothing. The only the dissimilarity is that the current threat is invisible. But it's still global, more

important than anything we have faced in our human history, and most people who have bonuses to pay in the next three months or whatever, are not interested, what's going to happen, then you're stuck. And that's how we humans work and that's why probably most other civilizations in the university they existed, followed the Fermi Paradox. When they reached the threshold of technology that can annihilate them, that's what they did.

Right. That's - for the benefit of anyone listening who's not familiar with Fermi's Paradox, Enrico Fermi, in about the 1950s said, "Where are the aliens? Because if life looks like it's so likely to develop on worlds, and there are so many worlds in the galaxy, then there has been plenty of time for many of them to develop to the same level that we have, why haven't we seen them? Why haven't we heard from them?" And there's no good explanation of that. But one of the more likely ones is that when they reach the point that we have where you might start to hear from them that they destroy themselves or something else takes them out.

That's the end of the first half of the interview. The remainder of it will be aired next week. I think this was a great illustration of how a futurist thinks and what they think about. The Ameca robot that Tony referred to, spelled A-M-E-C-A, is by a UK company called Engineered Arts, and it's the most lifelike humanoid robot I've yet seen. You can Google and find videos of it. Its facial expressions are remarkable and overall it looks a bit like the Sonny robot from the *I, Robot* movie. And it does speech recognition and conversation with GPT-3. Watch the videos, and it's easy to imagine it turning into the robot that you have a chat in the park with that Mark Lee was talking about in episodes 124 and 125.

OK, I don't do this very often, but here is my periodic reminder to you to like this show – check the Like box on social media in other words - write a review, whatever your podcast platform allows you to do, and share about it to your friends … or your enemies , for that matter, we're not picky. We don't do any advertising yet, and pretty much the only way that people are going to find out about this show is either if you tell them about it, or if the ratings climb high enough to make it more visible to them. I'm certain that there are many thousands of people, probably hundreds of thousands of people who would enjoy listening to the show as much as you do. So giving us a five-star rating is how you can help them find out about it.

In today's news ripped from the headlines about AI, *Wired* reports that one of Elon Musk's first actions on acquiring Twitter was to fire nearly the whole AI Ethics team. Rumman Chowdhury, head of Twitter's ML Ethics, Transparency, and Accountability tea, tweeted on November 3 that she was let go, saying, "Imagine feeling relief that you're the one on the receiving end of the Thanos snap." One other person said that all except one member had been fired, while another one said the whole team had been let go. That team had been doing research on political bias at the time. The work of this team was described by manager Joan Deitchman as pushing for algorithmic transparency and algorithmic choice, inventing and building ethical AI tooling and methodologies.

I hardly need to comment on this. If you've listened to any of our many episodes about AI ethics, if you've heard me talking about disinformation, you already know what I think of this. Twitter had been transparent enough to allow that team to publish the details of a bias it discovered in how photo cropping favored white faces and also a discovery that right-leaning news sources were promoted more than left-leaning ones. Anyone from that team who wants to come on this podcast, here is your standing invitation.

Next week, we will be concluding the interview with Tony Czarnecki when he'll be talking about what to expect by 2030 and just how that transhuman evolution into post humans will come about and what it will look like. That's next week, on *AI and You*.

Until then, remember: no matter how much computers learn how to do, it's how we come together as *humans* that matters.

http://aiandyou.net