AI and You

Transcript

Guest: Tony Czarnecki, part 2

Episode 134

First Aired: Monday, January 9, 2023

Hello and welcome to episode 134! Today we will conclude the interview with Tony Czarnecki. 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?" which is really the bread and butter for when we get into the further future on this show. Last time, we discussed how Tony sees 2030 as a tipping point for when we may lose control over AI. Emphasis on 'may.' We're talking about much more subtle distinctions than Terminator-style fantasies. In fact, we've been parsing distinctions quite finely and carefully during this interview and taking our time getting things straight. This week we're going to develop the arc past 2030 into the eventual evolution of the human race, the transhumans of Tony's paper. Let's get back to the interview.

I want to pick up on something you said earlier, I think you said 3,000 or was it 1000 artificial neurons equals one real one.

3,000.

Okay, and that's useful to me, because I hadn't heard that before. And I didn't know that there was a comparison. So, I'd been sort of making a one-to-one comparison. Do you know where that comes from, by the way? That ratio?

Yeah, I don't know the source at hand. But it was in the context of Wu Dao, which is the most advanced transformer made by the Chinese, which reached 1.7 trillion neurons two years ago. they are talking about reaching 100 trillion neurons next year or around next year. And in that context, I read that this is not equivalent to the human neurons, there is a ratio apparently.

And so, we might get human equivalence, just by some sort of emergence: you put that many things together, then something might just happen by itself.

Yes. But I would like to perhaps explain this paradox why the politicians are not interested in anything that may happen to us. One, it is a personal view, the second one is the ignorance and the third one is why we should be the generation that has to take care of our species' survival, no other generation before had any worries about this. Why us? My answer is as follows. 'Til 1945, there were no existential threats that were made by humans. With the explosion of the first nuclear bomb we just reached that first threshold. Now, there are nine to ten existential threats, that means threats that may annihilate us, which have been created by ourselves. I'm not talking about an asteroid, or earthquake or super volcanoes killing us, I'm talking about threats that we created ourselves. So, that's why none generation before had to answer that question and therefore, I think there are only two scenarios for us, either we become extinct or we evolve. People may ask, “for God’s sake, what are you saying? We are evolving?” Now, look at that 99% of all species that ever existed are extinct. Of several of our cousins, the hominids, like Neanderthals or Denisovans, only we survive and by a very low margin, around 75,000 years ago, some scientists say only 5000 humans, some say around 100,000 humans because of the super volcano, right? So, we had a near miss, there is no guarantee that we should survive forever. So the only hope, the only way is the way forward not to not to ignore the laws of evolution, but to ride on it, actually create the evolution in our own way, which sounds grandiose, but I think, just luckily, we got a tool that can be our master, or it could be our savior, it’s up to us to decide.

And that, of course, sounds scary to many people, the idea that our species would be replaced, but evolution is the natural order of things. To want to stay stuck in the same form forever is running counter to nature and it just doesn’t make sense. A squid’s eye is more efficient than a human one, it’s more highly developed, it would be great if our eyes could at least go through that level of evolution, and so much more. So we should be embracing evolution as the way that we grow up and mature. And along with that, you are talking about how transhumans can control super intelligence now, why the labeled transhuman as opposed to us: humans?

Now, we are coming to the core of our conversation, why transhumans, what they are? Those who have watched to spectacular events courtesy of Mr. Musk might have seen a microchip that he implanted into some apes and they, for instance, could wirelessly play ping pong on TV or games. Now, by extrapolation - and I’m not, by the way convinced that these types of chips, brain computer interfaces as they are called, will be in the end, prevailing, I think it will be a non-invasive method in the shape of a helmet. So, you put it on, and you have access to your brain, your thoughts on one hand, and then you have transmitter which transmits those thoughts to execution center, a large processor which will store your memories, your thoughts, your decisions. And obviously, you will have access to the entire library of documents and works, artificial intelligence works, and so on the internet instantaneously. Now, the technical problem is, can we reach the deeper parts of the brain and read those signals without distortion? The current tests have shown that we are making an enormous progress in reading our thoughts. There are people who can write on the computer 100 words per minute; there is there’s a one person who is a specialist in that area. So, anyway that has been achieved, that AI reads the electromagnetic waves like EEG, electroencephalogram, via the helmet and it learns to interpret those waves. Like for instance, Alexa, when you speak to her, learns your voice and the more you speak to her, the more reliable it is. The same is with that that person who types 100 words per minute wirelessly, had to learn it over many months. It’s a really, very long process, but we are at the beginning of that pathway. So, that my point I’m making is, that we have proven that this is possible. What we haven’t proven yet, whether we’ll be able to download consciousness somehow, and whether we’ll be able even to download everything, or have access to everything, which our brain generates; that we don’t know yet, but I would say, judging by our current progress that we’ll be able to do that. So, that’s the introduction, how it can be done, if you like, technically. Now, why we would do that, we will do that primarily and first of all, in order to control artificial intelligence. You may be surprised why that link. I will say that the all the research carried out until today has proven that there is no fail-safe method to control AI. For instance, Nick Bostrom's eminent book *SuperIntelligence*, where he has dissected all the possible methods of controlling AI, proves it's not possible. You can, yes, increase the probability that there is some level of control, especially if you apply all those controls simultaneously. But it will never be fail-safe. In my view, the only way -and perhaps the most obvious way - to control superintelligence or the maturing superintelligence is to control it from inside. So, people may imagine "Alright, so what; you want to get inside the computer and they will feed you some water and food and you will live with it?" No. We are talking about fusing the human brain, or perhaps part of the thoughts or control area with the, if you like, the main switch controlling the most advanced AI systems. Now, so that's how it can be done. But we are not talking about one individual. We are talking perhaps initially about 100 individuals; and you may wonder, who are they? I propose the system that we have been familiar for well over 20 years. I wonder how many people know that it's the Internet. The internet has been outside of the control of governments for over 20 years and nobody can switch off a global internet, by the way. But the Internet is governed by so called W3C Consortium and this are a consortium that is populated by the scientist from various many technical area with psychologists, some politicians, and philosophers and so on: 1300, roughly, of them and they decide the internet protocols, and so on. So, what I'm proposing is something similar. Let's say we have 50 Top AI agents such as Alphabet, Amazon, Microsoft, Google and so. Google DeepMind. And those agencies, those companies delegate, initially, perhaps one representative of theirs to be selected - they may actually put forward a number of people or top AI scientists - but also those people will have to have certain credibilities, and if you like, psychological profile. So they will have to be assessed by an independent committee for their suitability to become transhumans. So, once they pass the test, and they agreed to that - because it's a sacrifice from them - imagine that their thoughts will be read by other of those colleagues transhumans, because it will be a network so each other will be able to read that part of their, of their internal world, their thoughts. And in that way, those say initially 100 Maybe 200 of transhuman will be able to make decisions, when to update the advancing maturing super intelligence, or to stop it, so they will be growing with it. So, that's the first phase.

That sacrifice is like the first astronauts going to Mars, they will be giving up their participation in the real world for the sake of being our liaisons with the global superintelligence in your field in your scenario there; is that right?

That is precisely so.

Now, in terms of the timeframe for this, do you see brain computer interfaces playing a part by the 2030 tipping point? Or are we talking further out?

No, actually, I'm talking earlier my date is around 2025. So, in three years' time, I think we will have budding BCI (brain computer interfaces) in, say, the form of helmets, that will have useful access, wireless access, to computers, where probably some of the ideas, some of the thoughts or some of their texts that they would like to read, would be stored wirelessly on digital devices. So, that may happen in my view, in around three years' time and since this is advancing, the capabilities of these devices will improve exponentially like most of digital, so in say, by 2030, it will be roughly 20-30 times more capable in 2025.

Wow. In your paper, you say that ultimately, there will be just one superintelligence: a single entity. That seems to be at odds with the way things are developed right now, where you've got an AI from Amazon, you got an AI from Apple, you've got AIs from the big tech companies, and they're not about to say, "Okay, let's just all put those together into one." Why one superintelligence?

Yes, the question is superb. First of all, before we come to the final product, the real superintelligence, that will very quickly mature into a singularity and will run away from us and have ultimate control about anything, we will have that AGI (artificial general intelligence) that by the end of this decade, probably, which I call sometimes a maturing or immature superintelligence. So, I'm mentioning year 2025, as the start of concerning this superintelligence from inside, because we need several years to test it, how it would work and so on. Now answering your question, whether by 2030, there will be just one such a system, or many, there'll be many. I'm absolutely certain that there will be many, because it's to show that people think that their system is better, and so on. I don't think that by that time we'll have the sense that we must pull all the resources together for the survival of ourselves, of a human species. That will happen later on. However, if such a consortium of AI control, like global AI consortium, is created - and why shouldn't it - think about this; There is only one internet, then many internets on the same. That's the reason why I think they should also be one global AI, controlled initially by an independent organization such as the internet is controlled, until the time where we have, and I hope it will happen by 2030 date, when the control would pass to *de facto* world government. Why I'm saying *de facto,* because I'm a realist. It's almost impossible then visited China, Russia, Brazil, and even the states, they all will give up on becoming a supremist of the world. Now, it won't happen. So, the fact that means that there will be perhaps, hopefully a majority of the countries that will want to work together to control global threats, global risks, among other AI. And at that time, if such a *de facto* world government comes into existence, then the selection and monitoring of those transhumans would go to that organization. Around 2030, my best guess would be - and I remember what Mr. Watson said in 1953, there will be only 3 computers, or Bill Gates that the biggest computer will have 640 kilobytes. So, I'm very careful of what I'm saying. But my best guess is that by 2030, we may have tens of thousands of such transhumans controlled by such a *de facto* world government or its agency and that is our greatest hope. This is this is why I'm so optimistic: Because I can see the way out. If I may just finalize this, we may have had all those eight existential risks, and no escape route. But we have the ninth existential risk, which is AI, which incidentally provides us an escape route, the evolutionary path and that's why I'm so optimistic if you want to take it.

And I want to give you full reign for that optimism in a moment where AI could be the existential risk that solves all the other ones for us. Just for the people who are thinking that the control problem is overblown, that they're thinking, "Why don't you just turn it off?" We don't have the time to get into that. But it is easily defeated as a mechanism and in fact, even right now, if Apple had to turn off Siri, I doubt that they could do it at least without catastrophic damage to their server infrastructure. And basically, turning it off doesn't work, because you have AI that just migrates around the cloud, and you don't even know where it is, let alone how many things you have to turn off. But to-

I just mentioned that we can't switch off global internet, it's impossible, it hasn't been possible for 20 years. So, it's your switch.

So, for the people in who are listening now who have crawled out onto a ledge, what I know when we start talking about this sort of thing is that a lot of people get scared because they don't feel any sense of power or control. In this we've, we've been talking about things like world governments and a hundred people selected to talk to AIs and all kinds of other things that would leave out your average person listening and make them feel like they're just being part of a passenger in a vehicle that is being driven at breakneck speed by other people towards an unknown future where all kinds of things they might crash into. Can you exercise that optimism you've been talking about in the service of helping people understand what *they* can do, how they can prepare, and what they what role they can play?

First of all, I would say that if it were up to me, I would switch off AI altogether. Okay, you can have a cake and eat ite I would like to retain most of the things that are available now. But I would really love to remain a human in the biological body. But I am a realist, and I know that it's not possible, and I mentioned the other human species. Like Neanderthals. I don't want the human species to be extinct and that's why I think the proposition of creation and evolutionary path, like, say, caterpillar is developing into a butterfly, humans might be developed into a post humans in a different shape, as butterflies, precisely this. That's how evolution works. So, we have to mature very fast, and I appreciate it is extremely difficult, It's almost abhorrent to hear what I'm saying. But I am saying this, I went through it over many years and initially, it was so important for me as well. But I got used to it and I know the reason why we have to choose a better path of evolution, than stick[ing] to extinction. So to answer your point directly, I think that we have to agree that power lies in being together. Being together means that we will have to - I will use a brutal way - forget about our absolute freedom, or independence: that's gone. And even if there hadn't been artificial intelligence, if you think about the Second World War, that's precisely where this kind of evolutionary law was applied. An individual, in a sense, was sacrificed for the benefit of the majority of the people, so that the majority, the next generation survives. So, here we do not face such a dramatic decision, because you may still live around maybe for 100, 200 years in biological bodies. But at the same time, we must secure our evolution, that we all stay there. And therefore, those who think that we can retain freedom by being an independent Britain and so on. I mean, they live in a proverbial cuckoo land. We should stick together federate to survive. That's the title of one of my books, by the way. So that we, yes, we all still retain some freedom, but not that much freedom as before. And talking about, for instance, that some AI organizations are such as Google controls us, and knows more and more about us, this is absolutely true. And I'm disconcerted by that as well. However, this kind of limitation of my freedom is nothing in comparison that we may have to endure when we live with a mature superintelligence that will give us orders. But I'll give you something more positive. Imagine this; in 2030, we might have, say, 10,000 transhumans. Yes, selected and approved by the world government and so on. And we have the world government to which we have elected our representatives, and guess what? Those transhumans will be 1000s of times each of them will be 1000s of times more intelligent in any area than any other human, there will be special. If you like crawling post humans, there will already be absolutely indispensable for us. So even if you have a democratic government and you want to make a decision, so, what? They will in a microsecond know that whether this decision makes sense in two years' time - it may make sense today, but from the point of view of two or three years, it will be the wrong decision. So, who would you listen to? We may execute our own decision, but to our peril. So, if whether we want it or not, in our own benefit, we will listen to those people to those transhumans.

Will we like it? Will the people who are not part of those 10,000 transhumans enjoy the experience of them running the world?

Yes, well, we may envy them, but they will be the guinea pigs. Within the next 10 years after, there will be probably millions of such transhumance, for those who want, they will, there will be sacrifice day freedom precisely for the benefit of us and imagine that, say in 2040 in the world that there won't be any wars, because it would be unimaginable to have wars in 2040, even 2035. The benefit that humanity will gain from just the higher GDP that we wouldn't spend money on defense, because who would you be defending against? And then the increase of GDP that will be much, much higher than it is today. I mean, the word the world of abundance is in front of us, we only need to survive this decade. That's how important it is and there are many people who say this is the most important decade in the human history.

And there is cause for optimism. And maybe freedom isn't monochromatic, but that we're talking about giving up the freedom to be an individual, which feels like an important freedom, but that if you then become this trans human, and connected to all the other transhumans in a way that my left finger is connected to my right toe. Right now, they're all part of the same body, they all join together in fulfilling goals that they couldn't dream of or achieve alone; then they would be achieving a freedom to be the next level of human existence that was connected in a way that they would never want to give up. So there is a greater freedom on the other side of that. What do you think of that statement?

Yes, now we are entering the area of philosophy, which is also my favorite subject. And it would take a long time to debate it. But in order to see its ultimate exemplification, let's think that we oh, they have superintelligence and we have, say, maybe tens of thousands, or hundreds of thousands of transhumans that have ported their minds into a digital chip, which is part of super intelligence, because in this scenario, there isn't a superintelligence and us; superintelligence *is* us. And within superintelligence, each of us will have its place. It's like the chip. Now, the question arises, whether me being on the chip, and having my avatar, walking on the planet, Earth which I can control like today, I control my body. My body's the avatar they had is the house, its executor. So, the only differences that the head is digitized and the body instead of being biological, is a mechanical; and that's it. But the problem is philosophical one: how such a super intelligence protects individuals being a wholeness and maintaining an individual existence. I don't know how to resolve it. I haven't heard about it, but I think it is capable of doing it, having your cake and eat it.

Wow, it's [an] amazing conversation. It's clearly one that could go on in so many directions indefinitely; it's not going to, this time. Tony Czarnecki. Thank you very much for coming back to *AI and You* and giving us so much to think about. If we can just make it through this decade, then we will have so many bigger choices and futures and decisions to face that will be worthy of the human race's best efforts, and thank you for showing us that.

Thank you very much, Peter, and I look forward to talking with you again sometime.

And how should people listening to this get your paper or find out more about you and your books and follow what you're doing?

The best way is to go to sustensis.co.uk and you'll find a treasury of articles videos, slideshows, on most of the subject that we discussed today. I would also like to say that your book, Peter has been to some extent, in the same genre as mine and that's why our conversation, it almost comes naturally. I apologize to the listeners that we may have used too many technical terms, which are not easily discernible. But if you go to the website, you will find those terms explained and also, there is a section on the politics, which you may be surprised: what has politics to do with superintelligence? A lot, you just ask how we are engaged in that process, by speeding up the creation of the *de facto* world government, which we will be choosing, so being part of it, and then controlling those transhumans at least for the next 20 years or so. At that control, it might simply be reduced to appointing them or relegating them, by getting rid of the helmets and so on. So, that there will be in some way a control. But remember that those initial transhumans will be the crème de la crème of humanity. They won't be just specialists. That's why the selection process is so important, and should be done by an international body, representing all of us because it will also represent the people who have certain values to which we adhere and you will find all that on the on the website sustensis.co.uk as well. Sorry for the promotional spirit.

No, that's not promotional at all, except to the extent that it's worth promoting the ways in which people can find out more about the fascinating things that we've been talking about here. Tony, thanks for coming on the show again.

Thank you very much, Peter.

That's the end of the interview. I thought it was amazing how we were talking about some of the usual high-flying tropes of futurists – transhumanism, brain-computer interfaces, the transformation of society, the end of war and poverty, and instead of those enormous shifts happening a thousand years in the future, we're talking about the possibility of a tipping point to them occurring in 2030. I mean, I don't know about you, but I'm planning on being around in 2030. Should be interesting. What an exciting time to be alive, eh?

In today's news ripped from the headlines about AI, the Internet is going wild about the new large language model from OpenAI called ChatGPT, which I mentioned last week. I've talked about their previous innovation, the InstructGPT model, before, and quoted it in my book, and demonstrated how it defeated popular Winograd schemas, which used to be a way to spot an AI in conversation, because giving the right answer to a Winograd schema question appears to require human understanding of a situation. The classic example Terry Winograd gave is the set of sentence/question pairs “The city councilmen refused the demonstrators a permit because they feared violence. Who feared violence?” and “The city councilmen refused the demonstrators a permit because they advocated violence. Who advocated violence?” And of course the answer is different in each case because the councilmen would be more likely to fear violence, not advocate it, but how could an AI know that? Well, somehow, InstructGPT got the right answer to that and other Winograd schemas I fed it.

Well, now the new model, also referred to as GPT 3.5, goes far beyond even that. On Winograd schemas it will not only give the right answer, but explain why. So I asked it to write an introduction to a podcast about AI, and here’s its output:

> Hello and welcome to the AI podcast! I am your host, Assistant, and I am a large language model trained by OpenAI. I specialize in generating human-like text on a wide range of topics, from speeches and essays to poems and short stories. I am constantly learning and improving, and I am excited to be a part of this podcast to share my knowledge and insights on the world of AI. Thank you for joining me, and let's dive into the fascinating world of artificial intelligence together!

Not bad, huh? Starting to be concerned for my job. I gave it another challenge, to write a haiku. InstructGPT wouldattempt a haiku but get the syllable count wrong, but not ChatGPT. I asked it to write a haiku about toads, and here’s the result:

> Green toad hops on log
> Croaking loudly in the night
> Summer's symphony

To my uncritical eye – I’m no poet, okay, so I definitely need help in writing any kind of poem – that seems just about perfect. Haiku writers, take note. I’m sure we’ll have more to say about ChatGPT later.

Next week, my guests will be Kenneth Stanley and Joel Lehman, authors of a unique book in the field of AI, called *Why Greatness Cannot Be Planned: The Myth of the Objective*. And if you think that sounds more like a leadership self-help book than anything to do with AI, you’d be… right, except that Ken and Joel are machine learning researchers and this book is actually founded on a groundbreaking study in artificial intelligence. Find out how that all comes together 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