AI and You

Transcript

The Single Most Important Question

Episode 24

First Aired: Monday, November 30, 2020

Welcome to episode 24. We've been blessed with many terrific guests on this show and there are many more lined up, but today's episode is going to be a little bit different. My only guest will be virtual, which is to say, referred to only in their absence. You'll see what I mean.

Now it's important to me that this show doesn't just set its agenda by what our wonderful guests have been doing and thinking about, but that we take the most productive steps we can to address the fundamental questions that our fantastic voiceover artist opens every episode with. And today's episode, therefore, will address the most important question around the existential peril – or promise – of AI. What is that, and who is our guest-in-absence? Let me set the stage for answering that.

In March 2019, or what feels like about 25 years ago now, I was on a speaking tour of Britain. Since I was born and raised there, I set aside time to visit family. They're all still there; I was the black sheep who couldn't sit still. There was still more time left over, so I contacted my old high school and offered to come and speak there. They took me up on that offer, so I went. The school, by the way, was Southend High School for Boys, which has actually been taking girls since the last year that I was there; but this isn't the time to get into British idiosyncracies of terminology and tradition.

Now, I really like talking to students in the high school and college age brackets. They get what I'm talking about, and I don't have to sugar-coat it. I can tell them the whole unadulterated truth about the possible dangers of AI, and unlike many older people, they're not paralyzed or depressed. I tell them, "Our generation is handing this problem over to you. You can solve it. You *must* solve it." Our job is not so much to get out of their way as it is to stop holding them back. Our educational and social institutions were designed to suppress the energy of our youth in combining their efforts to create change at precisely the time of their lives when it is at its zenith. Maybe too many people in history got burned by student rebellions.

The way kids – and I mean zero disrespect by that term – respond is what gives me the most optimism about our future. So if you've got a school or college for me to speak at, please get in touch with me, because I get my batteries charged the most from that kind of interaction.

So there I was speaking to about 400 students at Southend High and afterwards, I decamped to a classroom to spend more time in a workshop with a smaller set of students. Now, the teacher

who had arranged my visit there, the very generous and helpful Chris Foley, had asked how he should select the students for that workshop, and I said, well, AI is going to affect all of us – that was the basis for my larger talk, where I related AI to just about every subject that was taught at the school. I said I could connect computer science to theology in about 30 seconds. Well, he took me at face value and invited the students who were studying computer science and the ones who were studying religious instruction and brought them in, about 30 kids.

So I went into the topic in a bit more depth and then opened it up for questions. And the topic was our familiar one of laying out the future progress of AI through increasingly sophisticated levels of automation through artificial general intelligence to its logical successor of artificial superintelligence, and the ramifications of our coexisting with those superintelligences, the sorts of scenarios explored by Nick Bostrom, Roman Yampolskiy, and Eliezer Yudkowsky among others. But when it came to time for questions – well, teenagers, especially English ones, can need some prompting. You could hear crickets chirping. So I said, "Stop being polite. If you're sitting there thinking, 'This guy's full of it, he doesn't know what he's talking about,' then say so. Just say *why*. I'm an engineer, I'm used to that. You should see what an engineering design review is like. Everyone else is looking for nothing but ways you are wrong. You're not going to hurt my feelings."

And this is where one of the students took me up on that. And he is our virtual guest, because I'm basing the rest of the episode around the questions he asked. I'll call him Jamie, because… that's his name. He said, "You haven't given us any basis for *when* all of this will happen." He was talking about the development of artificial *general* intelligence, which is what triggers the existential threats, because *super*intelligence becomes inevitable at that point. He said, and I paraphrase, "Do you have any data or model that would tell us when that is even likely to happen? Because without that, all of this is speculation that we can't do anything with."

Now, I *really* appreciated that question. Or I wouldn't be talking about it a year and a half later. But I started answering it by saying, "You're not going to be satisfied by my answer." I could indulge in that bit of mind-reading because Jamie was asking *exactly* the question I would have asked when I was in school. I knew just where he was coming from. And yet 18 months later, my answer is still not going to satisfy him. It still boils down to, "We don't know."

Okay, maybe you're thinking, "I guess that's the end of the episode; thanks for wasting my time." Well no, it's not the end of the episode and I very much hope it's not going to waste your time. Just because we can't answer that question doesn't mean we should shrug and ignore it. It's the most important question we can ask about our future with AI right now. It's like – imagine that that question is a giant wall in the middle of the desert extending to infinity in both directions and when we look up we can't see the top. Yet we really want to get to the other side. Do we just give up and sit down and draw circles in the sand? Let's test that wall. Let's see what it's made of, let's dig around it, let's see if we can start climbing it, let's move back to see if we can triangulate the top, let's do all these things that tell us more about the wall and maybe we'll learn something new.

Ironically, when I first started speaking about AI, being an engineer, I assumed that *everyone* would have Jamie's question. And while I knew I couldn't answer it satisfactorily, I constructed reams of evidence and arguments justifying my case. But what I found out immediately was that audiences didn't want that. Instead, after about ten minutes they would say, "We're on board with this. You don't have to convince us any more. What do we do about it?" And that question threw me for a loop, because I'd been assuming that once they became convinced of the need for action, they would figure out by themselves what to do about it in their world. I mean, how could I know what they should do? It's a different answer for everyone. But that's what they wanted, so I directed my energy to figuring out what I could do to help with that.

But here was Jamie asking that very question that I'd been obsessed with to begin with. You'd think I'd have been better prepared, right? But it wasn't so much me that wasn't prepared as the whole human race. No one else has a better answer either.

So what can we do around this question of **when** we will have artificial general intelligence? For a start, we can get better at defining what the heck AGI is. In 2017 there were 45 projects that stated they had AGI as their goal, so obviously a lot of people think they have a working definition of AGI, but in truth many of them may just be artificial *narrow* intelligence with a bow on top. GPT-3, which we've talked about on the show plenty of times, exhibits conversational abilities that ten years ago we would have unquestionably said belonged to an AGI, yet now we're a bit pickier, and we say, "Yeah, it can hold a pretty good conversation but it still doesn't have various other traits of an AGI that we'd figured would have come with those abilities."

Second, it's scandalous that our predictions of the rate of technological progress in this area are no better than voodoo. In that respect we're still just banging rocks together. Sure, we can predict *hardware* development progress with great accuracy through Moore's Law and its extensions, but that doesn't help. Even in the case of Moore's Law, which says that every 18 months or so we double the density of transistors we can etch on a chip – why is that? Yes, it's held true since 1970, but how come? If there were some generative principle we could point to, like: We're constructing manipulators that can build and use smaller manipulators, like in Robert Heinlein's story *Waldo*, well then, we could explain Moore's Law, but since that's not how we do it, something else is causing the progression and just because it's so reliably predictable, we take it as holy writ that it's going to continue. Now, I do believe that Moore's Law is going to continue – actually, the most important principle is not transistor density, because that would only impact the amount of silicon we needed, but what's more important is the cost of computing power, and that too is dropping exponentially. Which means that we can afford more and more computing. We can predict that trend quite reliably, and see how it will continue regardless of whether we continue to do it with silicon or some other computing substrate; but that is not enough to predict when we will have AGI, because we have no idea how much computing power it takes for AGI.

My opinion is that we *already* have enough computing power for AGI, because we can run millions of processors in parallel right now easily and cheaply. That's not something that would

take years or months of manufacturing and installing, but more like minutes or seconds of spinning up blades in data centers and configuring them to talk to each other.

When you look at how much computing power can be devoted to a single task right now, the numbers certainly stack up well against numbers like: how many neurons there are in the human brain. In other words, it seems very likely that we already have the hardware for AGI, and we just don't have the software.

There's another example of this inequality if you look at the cost of sequencing the human genome. When they first started that around 2001, the cost was effectively a hundred million dollars per genome. Now it's about $100. For a while, that cost fell according to an exponential curve like Moore's Law, because the sequencing hardware was semiconductor-based, but then it fell much faster, and the reason for that was not better hardware but better *software*, better algorithms.

The problem is, we can't predict when breakthroughs will happen. We don't even know how many breakthroughs or what kind of breakthroughs it will take to get to AGI.

You start to see why I told Jamie that he wasn't going to be satisfied by the answer; because *I'm not*. Does the fact that we can't predict when AGI will happen mean that we shouldn't be concerned about it now; and by concerned, I mean, spend our energy and attention on it?

Well, that's not satisfactory either. There's no law that says that we are guaranteed to get enough reliable forewarning of a threat to have time to construct a response to it. Suppose by the time we can prove that AGI is two years away we realize that it will take ten years to prepare for its effects? We'll wish we'd started earlier; but it'll be too late. That thinking is certainly what's driving a lot of people to spend time on this right now.

This is the point where I should mention the *Precautionary Principle*. I referred to that in the interview with Roman Yampolskiy. It's a very simple principle that's an application of statistical expectation, the sort of thing that underwriters and actuaries use . It says that if an outcome has a very low probability, you should still be concerned about it if it has a *very* high cost. So for instance, the chance of an asteroid hitting the Earth is very low. It happens only about once every fifty million years. (Yes, we're overdue.) But if it happened, it would wipe out *all* humans, so it has a very high cost. That means we should exercise *some* concern about it. And in fact we have. We've developed systems for scanning the skies looking for wayward asteroids. We've developed technology for moving them. (By the way, Hollywood's answer of blowing them up with atomic bombs is not the way to do it.)

Okay, can we apply the Precautionary Principle to AGI? If AI could wipe out the whole human race and it could do that some time in the next fifty million years, then we ought to spend at least as much money preventing that as we do on avoiding asteroid impacts, yes?

The problem is that we have no idea whether it's even on the scale of once in every fifty million years. The Precautionary Principle in the wrong hands is a dangerous weapon. You could use it

to argue that we should prepare ourselves to defend against an invasion of leprechauns from another dimension, or a zombie war, because both of those would wipe us all out too. But the problem is that we are very bad at dealing with big numbers. Intuitively we are about the same as certain cultures like the Walpiri tribe, whose numbering system goes one, two, many. We intuit a million as being much closer to a billion than one is to a thousand. A million years feels only a bit smaller than a billion years. However, it is in fact likely to be much longer than fifty million years between invasions of transdimensional leprechauns. Unfortunately, we can't rely on statistical methods to gauge how often AGI evolves, because it's something that only happens once and it hasn't happened yet. It would be nice to have examples of other civilizations where it has happened, which is why we listen for signals from other civilizations in the galaxy. Unfortunately, we haven't heard any yet. Bummer.

By the way, you may feel that I've taken unwarranted license by simply assuming that the development of AGI will inevitably lead to artificial superintelligence, which is where the real existential threat lies. Shouldn't I have to prove that? No, I don't, not for Jamie's question, because AGI is a hard enough question by itself. A lot of research suggests that artificial superintelligence – ASI – is not just *easy* as soon as we have AGI, but unavoidable. But AGI alone – the development of intelligence even only on a par with a human five-year-old, but able to understand everything that a five-year-old can – would require fundamental breakthroughs. Or so I and others think. There's a small but growing cadre of people who think that our current deep learning models will manifest AGI when they get enough parameters. How many might we need? Perhaps as many as there are connections between neurons in the human brain, which is around a hundred trillion. GPT-3 has 175 billion parameters, which puts it about three-hundredths of the way there. Which is not that far off, considering that its predecessor, GPT-2, had a hundredth of its parameters, and that was, what? – only a year earlier. But to Jamie's frustration and mine, they can't prove that developing a learning model with that many parameters will exhibit artificial general intelligence any more than I can prove that it can't.

Now, I could say that more and more people; brilliant people, experts, around the world are becoming concerned about the potential evolution of AGI and think that it's worth putting our attention on it now. I'm sure Jamie would be sharp enough to point out this argument as an appeal to authority; but there again, sometimes we need to depend on authority. At some point, I'm not going to personally verify the Law of Gravity, I'm going to take Isaac Newton's word for it, because I don't have time to measure it myself. But there again, it's not just Newton that I'm trusting, but the numerous other scientists who didn't take his word for it but *did* personally verify it. Does that mean that some critical mass of researchers being concerned about AGI would be enough? No, not by itself, not unless they were conducting independently verifiable experiments published in peer-reviewed papers.

It's not enough to say, in other words, that a whole lot of smart people having a bad feeling about this can prove anything.

It is possible, after all, that the people who are the most concerned with this threat are smart enough to be able to see further into the future than most of us – or if nothing else, they simply indulge in looking into the future more than the rest of us – and so they're warning us about something that is going to be a threat on a time scale far enough away to be of little concern. Science fiction writers routinely think about things that may not be feasible for a thousand years while the rest of us get on with our lives. That's possible, too. But it's also possible that fundamental breakthroughs in cognitive algorithms could happen within five years, and precipitate all of this then. We just don't know.

This question seems to resonate with the old adage that "If you can't measure it, you can't manage it." Which I have long declared to be BS. It's a refuge for managers who won't face hard questions to hide behind. But some of the most important factors in leading employees, like loyalty, friendship, and passion, are unmeasurable by useful means. People join or leave companies because of intangible reasons like whether they were challenged, whether they got along with the other people there, or whether their boss listened to them. So I don't agree with that adage. It would be nice, of course, to be able to put a date on when we'll get AGI. But that doesn't seem to be something we're likely to be able to do.

Some authorities *do* think they can put a date on this. Ray Kurzweil says that we will have human-equivalent computing by 2029, and the Singularity will happen in 2045. But his basis for those dates is shallow, and not shared by many independent researchers, although he has plenty of followers who want to believe in those dates. Largely, those predictions are based on estimates of when we will have computing power equivalent to the human brain. The big problem with that is that we have little idea of what the computing power of the human brain really is, and comparing it with the instruction processing functions of silicon computer chips looks very much like an apples-to-aardvarks comparison. Our computers are based upon the von Neumann architecture of a processing chip, a separate memory bank, and input/output interfaces. The human brain has almost no resemblance to that, and looks like a more or less homogenous sponge of neural networks.

Frustrated yet? Join the club. That is what it means to be involved in this field. Yet, as I said, we can't just say that our inability to put a model with numbers behind our predictions means that there's nothing to worry about. We can certainly demonstrate that a breakthrough in AGI development could leave us insufficiently prepared to develop an adequate response.

This is where we have to exercise lateral thinking. Instead of tackling the question head-on, we must think at right angles to it. We could ask, "How could we defend ourselves against artificial superintelligence?" That question has been addressed quite thoroughly and the answers are, shall we say, disappointing. We could also ask, "How can we *prevent* the development of artificial superintelligence?" and that question has also been asked and answered to some extent.

Just as an aside, by the way, there is a credible argument for saying that we should develop artificial superintelligence as soon as possible. And this is one of those arguments that sounds ludicrous to begin with, but when I explain it, it will make perfect sense. The argument for doing that is that we should get our practice at dealing with artificial superintelligence when the computing hardware is at its slowest, so that that superintelligence evolves as slow as it ever could, and gives us the most time to deal with that evolution.

In other words if we wait until it happens later, then our computing hardware through Moore's Law and it's extensions will be going that much faster and the artificial superintelligence will therefore evolve on us on us that much faster at that point. So getting it as soon as possible gives us more time to deal with it. Counter-intuitive? But it makes sense.

Anyway, one useful question would be, "What kind of technology would give us the ability to *predict* the arrival of AGI?" Maybe that question is just as hard to answer, I don't know; but it is a *different* question.

We could ask, "How can we make our social and governance structures more resilient to the emergence of AGI?" and that would be useful, regardless of any existential threats.

Likewise, we could ask, How we can adapt our educational and research institutions to direct more energy to these questions. We have spent comparatively little energy on them so far, and it would be silly to think that the big question is unanswerable just because our first glance at it is on a par with a dog looking at the blackboard in Calculus 101. We have lots of room to get smarter. Who's a good boy?

So Jamie, I'm sorry we're not closer to the answer you wanted. If the answer were proportional to computing hardware speed then it would be easy; even if we were only one millionth of the way there, in that case, we'd reach the goal in 30 years. That's what our improvements in computing hardware will buy us. That's a complete guess, of course, but at least note that 30 years would be within your horizon of concern even if it isn't for some of the other people listening to this podcast. But the answer may have very little to do with hardware speed, and everything to do with algorithms, in which case we get into the commonly-cited territory of wild estimates ranging from 5 to 500 years.

But in a way I'm also not sorry that we don't have an answer, because as I said in the first episode, questions have more power than answers. My books, m y talks, this podcast, are all about asking good questions that sit with you and cause you to ask more questions by yourself. Questions drive us forward.

Alright, that's enough excavating around this topic for now. You may or may not see how this is immediately applicable in your life or what you can do about it right now; but the point is to raise awareness of the important issues so that they have the opportunity to percolate in the recesses of your brain where they can create inspiration at opportune moments. After all, there are a great many more people working on this problem now than there were ten years ago, and

they got their inspiration to do that from somewhere. Maybe this podcast will be your inspiration.

In news from today's AI headlines, humans aren't the only ones to have had their travel plans ruined by the coronavirus. A robot-powered boat that was due to cross the Atlantic has been forced to delay its voyage until next April, after the virus caused complications in its development.

The autonomous 15-meter trimaran has been built to push the boundaries of autonomous shipping while gathering scientific data on the ocean. The *Mayflower Autonomous Ship* (MAS for short) is being led by marine research organization ProMare, and IBM is the main technology partner.

The solar-powered vessel was launched on Sep. 16, the 400th anniversary of the *Mayflower* departure in 1620. It will go on several voyages and missions over the next six months ahead of a transatlantic voyage in April 2021.

During that transatlantic crossing, the ultramodern ship will broadly retrace the *Mayflower's* original route from Plymouth to Cape Cod's Provincetown. IBM says, "The portal even features a seven-armed, stowaway octopus chatbot called Artie, who claims to be hitching a ride on the ship. Powered by IBM Watson Assistant technology – what else? – and created in partnership with European start-up Chatbotbay, Artie has been trained to provide information about MAS and its adventures in a lively, and accessible format."

In next week's episode, we will talk with Thomas Homer-Dixon. He is the director of the new Cascade Institute at Royal Roads University in British Columbia. Between 2009 and 2014, he was founding director of the Waterloo Institute for Complexity and Innovation. He has a PhD from M.I.T. in international relations, defense and arms control policy, and conflict theory. His latest book, *Commanding Hope*, explores how we develop and maintain positive and useful outlooks on our existential threats, whether ecological, environmental, or technological. The Cascade Institute researches and models the global systems that drive crises and our responses to them, looking for intervention points. Those crises of course include technology, including exponential technologies such as artificial intelligence. 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