AI and You

Transcript

Special Episode:

Episode 104

First Aired: Monday, June 13, 2022

Hello and welcome to Episode 104. This is our two-year anniversary at this point. And I just want to thank you all for hanging in here with this show throughout the pandemic. That's where we started out: so small; you can see when there were just a handful of people listening to this, the first few episodes. And now with no advertising whatsoever, we have a thousand listeners an episode. Imagine being in an auditorium with a thousand people around you listening to a talk that's a pretty big crowd, that's the size of the TEDx hall I last spoke to. That's who's listening to this with you now. Only instead of being under one roof, you're all around the world and I hope you're doing well because we've had so many challenges. There's been such a cost of human lives and suffering throughout this pandemic. I'm really ready for it to be over. And now we've got the war in Ukraine and the economic situation with inflation, we deserve a break. And so anyway, thank you for listening. And if you want more people under this roof listening to this show so that we could get some community together going and participate with each other in other ways, then share about it. Like it, give it five stars, say something about it that's nice and then they'll find out about it. There are a lot of people who would love to listen to this show. They just don't know about it, I guarantee. And your saying something about it or giving it a good rating, will help them find it.

So in this show we're going to be talking about some of the terms around different types and levels of advanced artificial intelligence that are of interest to people who are tracking it from a broad historical or philosophical perspective, terms that we've used in the show before and not always explained or never explained and now we're going to do that. So no interview this week, just me talking about terms for different levels of artificial intelligence to begin with. You may have heard us use terms on the show like AGI, ASI, and ANI. What do those mean? ANI stands for artificial narrow intelligence which is the AI that we have. Now every kind of AI that you know of or have met is artificial narrow intelligence which is to say it does one thing, but extraordinarily well. Like the UNIX philosophy, if you know that one. Usually better than people, otherwise there wouldn't be much point in doing it. It could be interpreting a radiological scan, it could be facial recognition, it could be playing Go or chess, it could be running a nuclear fusion power plant. It could be many different things. But it's only going to do one of those things at a time. Now there are signs of development coming out of Google at the moment that could be able to do several different things, possibly even hundreds of different things and learn how to do those at once, which is a very significant achievement. But the number of different tasks that human beings can do is probably in the millions if not billions, I don't know if anyone's tried to count that. And so even if we had somehow trained different AIs to do all of those millions or billions of different things, the big problem still unsolved would be, how would an AI decide which one of those to switch to in the right situation? That would be a huge challenge. That would be a Nobel Prize right there. So what sort of thing could do that? And that is the next type of AI in the scale, which is artificial general intelligence or AGI. Doesn't exist yet. But a lot of people are working on it and some of them would like to say that they're already there. So you will see advertising that says AGI, we're doing AGI.

Now this would be artificial intelligence that can solve general problems like people do, even if only at the level of, say, a nine year old person. In fact, if we could do even what a human five year old does today, particularly in the embodied application of acting within the world and physically manipulating things, we would be far ahead of the state of today's art, in solving general problems with AI. Now this is where it's important to not get caught up in thinking about how Ai is able to do some things that are clearly superior to what humans that are nine years old or any age can do. Because those AIs can only do one of those things at a time. So an AI might be the best chess player on the planet but that program is not going to do anything else. It might be possible in the case of say Alpha Zero to have an AI that you can train to play go and then you can later on train the same engine to play chess. But still it's only doing one at a time. Whereas at the moment we really don't have anything that's capable of solving the wide range of problems that a human five year old can: find a toy on the shelf and take it down, take it apart, put it back together, make a peanut butter sandwich, draw a picture of the family cat, put on a t shirt, tell you what happened in the show they just watched, pet a hamster, talk to the hamster. You know, I could obviously go on for ages, you get the point. Each one of those tasks would be a monumental achievement for today's AI.

Now the money currently being expended on researching AGI is in the many billions of dollars. I don't know if anyone has added it all up, but it might well be hundreds of billions of dollars. Why would it be so useful to develop something in a computer that was only as smart as a nine year old or five year old and nine year old? Well, let's leave child labor laws out of the picture. 18 year old humans are a lot cheaper in general than a billion dollar AI. Well, of course they are, even if it does take 18 years to make one. But if you had such a thing in AI it would be the equivalent of a nine year old person with all of Google and Wikipedia in their brain, able to access the world's information instantly and perform calculations at the rate that we're accustomed to seeing them done by digital computers, which means that they could run environmental simulations, protein folding calculations, and optimize the power consumption of a data center, for instance, in their head. And that of course would confer an enormous advantage to something with even only the general reasoning capability of a nine year old person.

Of course, some people have figured that it might be actually easier to put Google or Wikipedia in a person's brain than to build artificial intelligence that has the brain part down. And this is called augmented intelligence also spells AI. Doesn't seem to be confusing yet though. And the brain computer interface such as the one that Elon Musk's Neuralink is developing represent one possible way of getting there. So an artificial general intelligence would have an incredible advantage over non augmented human beings because of the speed with which it could manipulate data.

If you had an artificial general intelligence that was say running a company, a big one, think Walmart, IBM, google, it would be able to perform all of the roles of the current C suite: CTO, CEO, CFO,… but wouldn't have to wait for them to become available for meetings or arrive at decisions or gathering process information and handed around from person to person. It would simply do that instantly. That would mean that it could see a business need, analyze that from data, conduct studies and send a decision to a manufacturing line for a new design before the human-based competition had finished reading the quarterly sales report.

An army that had an artificial general intelligence in charge of it would be able to analyze a situation and devise a strategy countless times faster than humans in the same situation. To deal with either the army or the company in our world now it takes a number of humans working together at peak efficiency and

communicating with each other as frictionlessly as possible, and reams of books and papers and studies have been written about how to do that better because it's just about the biggest problem that any group of people working together face. And the reason that those things aren't done by one person who wouldn't have to wait for conversations to take place, meetings to resolve, and decisions to be made is that no one person can fit that much in their head. Of course. We have reached our limits. But there's no reason that any artificial intelligence should be so limited in that capability. In other words, you could have an AGI. That does the job of the CEO, CFO, CTO, COO. All of them put together. If you can do any one of them, there's no reason why it shouldn't do all of them and then not have to wait for them to coordinate with each other.

Now, perhaps you can see why so many people are working towards artificial general intelligence. Perhaps you can also see why Vladimir Putin said that the country that develops it will "rule the world." And boy, has that quote gotten tinged with more meaning in the last six months.

Now, despite the fact that the advent of AGI as I just discussed would be a historical game changer, it's not the end of artificial intelligence development. The phase after that is artificial super intelligence or ASI. I know it sounds like these things should actually be called narrow artificial intelligence, general artificial intelligence, NAI, GAI, and so forth, but this is what stuck. So here's the terms. Now, ASI would be AI that was generally intelligent still, but many times more so than us. Think of how inadequate you feel compared to Sherlock Holmes say, if that's something that's within the scope of your imagination or pick Einstein or someone that's similarly at the peak, the zenith, of human intellectual capabilities and then multiply that by a factor of a million and you would get what Einstein would be feeling like next to an ASI. Now we're out in speculative realms here of course. But as I've said on a previous episode (#24), we just don't know and can't predict how far that is to any useful level of accuracy. I'm not saying that means it could show up tomorrow, I'm not saying it means it could show up in five years from now, but it is also true that no matter how far out it is, it will take an extraordinarily long time of coordinated effort to figure out how to deal with it when it arrives. But that's another conversation, go listen to episode 24.

Artificial super intelligence started coming in for a great amount of popular scrutiny when Nick Bostrom wrote the book Superintelligence in 2014 and people like Elon Musk in particular devoured it and took it as a call to action. Now it's important to note that Bostrom is not a computer scientist, he's a philosopher; he was not concerned with when super intelligence would happen or even whether it was possible at all, but only with researching what would be the consequences if it ever *did* happen. And it's fair to say from his book that those consequences are pretty bad. This is not the time for me to go into those consequences or what we should do about them. But it's worth noting that a great amount of attention is focused now on artificial superintelligence. As a result, as I have said on other occasions, talking about the existential risk of artificial super intelligence is a conversational black hole. Once we venture into it there is no going out. We will see whether that's possible on this episode. Though a great deal of debate has ensued over how likely we are to get artificial super intelligence. But it's actually, in my opinion inevitable once we get to artificial general intelligence because anything that has reached the AGI stage of embedding the human ability to learn from its environment into silicon or some other kind of chip has already started on a path that it's capable of following much further than humans do because it would not have our biological limitations. Which in our case means it would not be limited by the size of head that a human female can fit through her birth canal. Whereas AI is effectively unbounded.

We've already demonstrated one of the ways in which AGI might become ASI. Which is recursive self-improvement. An AI that learns how to get better at learning. Because this is what we have done. What is that? Because this is what DeepMind did with AlphaZero, which was their deep learning engine that taught itself to be the best Go player and the best chess player on the planet within hours just by playing against itself. So as much as people, some people, might like to think that if we created AGI, we could constrain it to be at a certain level of sophistication and not go beyond that, it's almost certain that it will eventually encounter a situation where it will learn faster and faster. Or that someone will deliberately program an AI to do that. That's what we call the "hard takeoff" in these circles.

So this brings me to a couple of scenarios that are discussed heavily in the AI existential risk circles which those people call X-risk for short. The first one is the Control Problem. In other words, how do we ensure that we can maintain control of an AI as it becomes more powerful and more capable? Now as much as this might sound like it is coming out of thinking about Terminator-like scenarios - and that's certainly the label that the press slaps on it as soon as they hear this kind of discussion - there's actually a lot more to this scenario that has nothing to do with killer robots or any kind of self-aware or conscious. An AI that becomes more capable and complex will have failure modes that are equally high level and it's the goal of virtually everyone working in the artificial general intelligence today to create an AI that is capable of responding to general instructions and taking initiative to figure out how to follow them. That's what human beings do.

However, it's very difficult to specify where the boundaries of initiative should lie when human beings give instructions to other human beings. We assume a mutual context that includes standards like "don't kill people in the pursuit of this goal." Of course, there are exceptions to that rule, but they're pretty obvious and we know where they are. Businesses are notorious for overstepping the boundaries of what's legal, ethical or moral and considering what they can get away with. But even in the most heinous of cases, human beings are constrained by their abilities. Now, an artificial general intelligence might be able to crack any kind of computer security and get into any kind of system and thereby control anything that's under the control of a computer in pursuit of a goal, and anyone who is familiar with the current state of cybersecurity would understand that we shouldn't have any confidence that we could keep out an attacker that had literally firsthand knowledge of computer architecture.

So an artificial general intelligence or an artificial super intelligence might have the capability, whether we intended it or not, to control any physical system that's reachable by a networked computer device. And by extension, reach any other system that is reachable by a device that can be controlled by a network computer device. You can see therefore why people are thinking already about how to ensure that we can control such an AI. This isn't the time for me to go into the research that has been done on the Control Problem. But I will just point out that is not yet solved.

A couple of researchers that I'll mention here that are working on the control problem, Roman Yampolskiy who has been on the podcast and Eliezer Yudkowsky, and they're looking at how can we ensure that an AI cannot change the goals that it has been given. But even being given certain goals, we still have the problem with it pursuing those goals in the wrong way. And this is framed as the King Midas problem or the Genie problem. You only have to look at the story of Aladdin to realize that if you ask a genie that grants you three wishes the wrong question… well, you know how this always turns out. The first wish is for something magnificent, incredible, could be greedy, could be altruistic. The second wish is trying to undo the damage done by the first wish and the third wish is to wish for everything to

go back the way that it was. Likewise, King Midas thought, wouldn't it be great if everything I touched turned to gold, but didn't realize that that would include food and children.

And this problem is raised in particular in another trope from Nick Bostrom, the *Paperclip Maximizer.* The scenario has gone down in legend now. So people just refer to paperclip maximizing when they want to talk about this problem. And it's one where the scenario is that an AI is put in charge of a paperclip factory and told maximize paperclip production. It's the sort of direction that a CEO would give to a COO and just let them figure out the rest. But the AI does not have in this scenario the huge context of what's permissible and aligned with human values, and so goes off and says, well I see a lot of things outside this factory that are not yet turned into paper clips. The fact that those are cars or pets or people is irrelevant. As Yudkowsky puts it, the AI does not hate you, it does not love you, but you are made of atoms that it can use for something else. And so in this scenario with a suitably capable artificial general intelligence or super intelligence, it ends up consuming the entire world to turn it into paper clips because it doesn't know any better. And even if it didn't have the control necessary to do that in the first place, it's still motivated by its goal to do whatever it can to acquire it.

Now, another problem that's related to this is called the AI Boxing Problem. And it was visited especially by Eliezer Yudkowsky, and that's the question of can we keep an AI in a box? It's sort of like can you keep the genie in the lamp, and only let it out to grant your wish. In this scenario the idea is that we keep the ASI inside some isolated environment where it can't get out and control anything else. Air gapped to the max, lead-lined room, independent power source etcetera. And that seems pretty foolproof. But Yudkowsky decided to test this. Now of course we don't have an artificial super intelligence to do that with. So he played the part of that. That was obviously conservative because he would admit that he is not an artificial super intelligence. And he tried to see whether he could talk someone acting as a gatekeeper in a hypothetical situation into letting him out of the box, and succeeded three times out of five. It was a carefully controlled experiment. So you shouldn't think that that was something where any kind of cheating avenue that you can think of was exploited. And in fact Yudkowsky frames it as, this is seeing whether it is possible to control a human mind through a text only interface.

Now, you could argue, as I have, that perhaps we don't need to control AI. I have Children. I can assure you - and any parent knows this - that my ability to control them is diminishing by the day and any that I might have already will someday be gone altogether. Nevertheless, I am not concerned about them destroying the world, possibly not for the want of trying but also because they're not able to destroy the world. But there again - I think we might be on the third hand by now - that might not be true tomorrow. If you've read my first book, you know that part of it focuses on the problem of control of synthetic biology and the capability to construct viruses that are lethal and far more contagious than the ones that we have right now. Yes, even coronavirus. It is rapidly becoming within reach with technologies like CRISPR and DNA synthesizers which fit on a desktop.

So the problem of controlling an all-powerful AI might be the same as the problem of controlling all powerful, in other words, soon to be all of us, human beings. And maybe then what we should focus on is not controlling it, but ensuring that it doesn't *want* to destroy us or that it doesn't accidentally destroy us to give it values that value. The same thing that we do, including ourselves to make. In other words, friendly artificial intelligence and that's another goal that a number of researchers have. Of course, one

of the problems with that is as much as we might create all kinds of friendly ai. It only takes someone else creating an unfriendly Ai to upset the apple cart.

But this discussion about values brings us to the other big problem that I wanted to talk about which is called the Value Alignment problem and that is literally what it says on the lid. The problem of aligning the values of an AI with human values. Stuart Russell gives a classic example of an AI that ends up cooking the family cat because it's been given the goal of "serve the family." It knows that it should protect the family, knows it should feed the family and one day it sees that the refrigerator doesn't have what it needs to make dinner. But it does see the family cat and is aware that the cat is a source of protein and well you can figure out the rest. It wasn't given the value, someone forgot to program into it, the value that the family very much like the cat and that this is not something that it should do. And even if you think that's an obvious example to solve then if you make the cat a dog and you translate that to certain other cultures, you can see how the lines could get blurry.

Now, Stuart has been doing some really interesting work lately about how to make AI that doesn't accidentally run away with the wrong values. And broadly speaking, as I understand it, it means teaching the AI that it doesn't know what human preferences are and that it should test at every stage to see how well its assumptions are doing and back off and not do anything and find out more whenever it looks like those assumptions are not working out, which is a pretty good model for humans to follow too, by the way. Read his book *Human Compatible* to find out more about that.

Now of course we think that having values implies consciousness because everything we know that has values seems to be conscious right? But actually, everything that acts has values. The more complex the system, the more meaningful the values. It's easy, for instance, to impute values to an organization. The company's actions are more than the sum of its employees' actions. It could go through 50% or 100% turnover and still be acting in the same way because there's an organizational culture that is self-perpetuating, which lives on in the gaps between the people. It's an emergent value system. And some people have said that corporations, big ones, are already super intelligences, which in theory could be true. But often in practice, those organizations involve some degree, usually a high degree, of internal politics that greatly limits what those people are able to do collectively.

How should we specify human values? Of course, this makes people think of Asimov's Laws of Robotics which he enunciated back in, I think the 40s, where he said the First law a robot - for which we can read AI - A robot must not harm a human being or through inaction allow a human being to come to harm. Second Law, a robot must obey the orders given to it by a human being in all cases, except where this would conflict with the first Law. Third Law, a robot must protect its own existence except when this conflicts with the 1st and 2nd laws. Now that sounds pretty good. Asimov wrote these in reaction to the stories that had been coming out at the time, which were universally painting robots as monochromatic bad guys that would kidnap helpless females and carry them off for unspecified acts. But of course, those laws have all kinds of loopholes, and in fact that's what Asimov stories then did was explore all the loopholes he could find in those laws, so they're not an adequate starting point.

There's a vast background that needs to be made explicit for AI about what we consider acceptable, like not turning people into paperclips. But if you think about our own systems of values, we can't even align to each other's values and that's why we have wars. But we even can't align to our own values, the values that you have tomorrow morning at 6:30 a.m, when the value you have today, that says "I'm going to get up and hit the gym" may change by then - just a wild guess. But for many of us that

situation tomorrow is different from what we are anticipating today. And a less trite example would be someone going through detoxing and now the place that they have gone to has to consider who they will be and the values that they will have when they have finished that program.

So if we can't even figure out how to align with our own values, let alone other people's how is an AI supposed to do that? There's a problem for computer scientists who are used to having to specify things precisely in order to get a computer to do it. And so a lot of the thinking around the Value Alignment problem consists of trying to reduce human values to neat formulas that we can then express in a computer or determine whether or not the computer is meeting them.

Anyway, that's a tour of just a few of the terms that we've been using in these shows. And so now hopefully you're a little more informed about those when they come up in the future.

In today's news ripped from the headlines about AI, researchers at the Japan Advanced Institute of Science and Technology have integrated biological signals with machine learning methods to enable emotionally intelligent AI. Emotional intelligence could lead to more natural human-machine interactions, they say. Well, yes. There's that famous quote that 93% of communication is nonverbal which is almost an oxymoron because how much more verbal can you get than a number, especially one as precise as 93%. But it is certainly clear and undisputed that most of our communication with each other doesn't lie in the words, but how they are said, our facial expressions, and how we're gesturing. So you might wonder how did the researchers deal with that? And the answer is in their study in the IEEE Transactions on Affective Computing. Affective Computing is the term for emotionally intelligent AI. And it goes back to Rosalind Picard at MIT and Rana el Kaliouby at Cambridge. The researchers added physiological signals to the range of emotional state that they could detect from speech tone, facial expressions and posture, which are aspects of emotional state that affective computing was already measuring and are collectively called multimodal sentiment analysis. The physiological signals that they added were nonobservable factors such as skin potential, heart rate, breathing rate.

Now, Professor Shoho Okada said that humans are not brother now Professor Shogo Okada said that humans are very good at concealing their feelings, but the internal emotional state is more easily observed through those biological signals that I mentioned. And what they found was that they were able to train the AI to interpret those signals that are not normally observable by human beings to do a better job of estimating sentiment in this case to assess how someone found in a conversation enjoyable or boring. Yes, it's easy to interpret some sort of dystopian use of this technology, but it could also be used in for instance, emotional therapy. The researchers concluded by saying nevertheless, further research is needed to realize an emotionally intelligent agent with beyond human capabilities.

Next week, I'll be digging into our archives to bring you an interview I conducted at the 2016 Canadian Artificial Intelligence Association Annual Conference, never before aired, with Professor Michael Bowling, who created an AI that could provably win at poker. 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