AI and You

Transcript

Special Episode: ChatGPT

Episode 141

First Aired: Monday, February 27, 2023

Hello, there are a lot more people listening to this show now. No prizes for guessing why. It's all the buzz about ChatGPT that has focused the attention of literally millions more people on artificial intelligence. So welcome regular listeners and new listeners to episode 141 and welcome to *Artificial Intelligence and You*. You will find a wealth of over 100 episodes starting from mid-2020, all designed to help you make sense of this thing called artificial intelligence. And we will talk about ChatGPT and help you make sense of that. So first for you new listeners, what is this show about? Because I'm realizing that in some repositories, this show's first episode may have dropped out, may no longer be available, but if you want to get a good idea of what we are here for, then listen to episode number one.

So people have a lot of the same kind of questions about AI that are prompted especially by ChatGPT. "Should I be scared? Am I scared? What's going to happen to jobs? Is someone regulating this thing? What even is AI? What can it do? Can I use it?" So as I said, we are here to help you make sense of AI. It is multifaceted. It gets into all human fields, I mean all of them, from industry to government to education and all aspects of human expression from creativity to wellness to ethics. I know of nothing else like that in human endeavor that has so much application over everything we do and everything we are. There are so many ways to think about this and all metaphors are imperfect and incomplete. But one way of looking at it is that AI is a mirror for us. That we're growing a new species before our eyes in some ways in our own image and in others not. In some respects, it already exceeds our abilities of course, and in others is incredibly immature. But we don't know how it is going to expand those capabilities next and we are continually surprised by advances such as ChatGPT doing things that we thought were going to remain exclusively human for at least some time to come. And this is true even of the people who invented it, that they are surprised by those capabilities. But one thing to bear in mind is that what we are discovering with what ChatGPT does, is that it's more revealing of how *humans* work than it is of any general intelligence of the AI. More about that later.

A little bit about me: I have been helping people ask and answer those kind of questions about AI for over 10 years. So here's a little bit of horn-blowing: In 2017, I wrote a book that started with a prediction of a global pandemic evolving through multiple strains, each of them more troublesome than the last. And it also included an AI that you could have a conversation with, and which would pass the Turing Test, and AI that could write software to specification: you just tell it what you want and it will write the computer code for it. I know that chapter was fictional in that book, but I'll take a bow anyway. I thought I was going out on a limb at pegging those kind of advances for 10 years from then, which meant 2027. I would plead artistic license when I was talking about that part of it. And until a year ago, those sort of advances still looked optimistic. Now they look too conservative since obviously we had the pandemic and ChatGPT more or less fits the description of the AI that I was describing there.

The AI in the book was named Sybil. Not that it was explained, but because she presented many different faces. Just a bit of an in-joke there. And fiction aside, the kind of event that we have seen with

ChatGPT is the kind of transformative disruption that I and others have been predicting for some years. We have said that this is coming. We may not have known the time, the place, the exact nature, but we knew that at the rate that artificial intelligence was evolving, that something like this was going to happen. It's our job to look for those kind of things and you'll hear more later on in this episode about what might be coming up in the future. That book of course now nearly six years old and I came out with one last year that is much more current and has twice the content of that former book. So I would encourage you to get that; more about that later.

So if you don't know about ChatGPT; well, we will illustrate that in a number of ways. It's an online chatbot. It was released on November 30th. Hard to believe considering that this is now February. And the day after it was released - perhaps this illustrates its capability rapidly - a tweet from one Thomas Ptacek said, "I'm sorry, I simply cannot be cynical about a technology that can accomplish this." And he instructed it: "Write a biblical verse in the style of the King James Bible explaining how to remove a peanut butter sandwich from a VCR." The response was lengthy and started out, "It came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it." Then it goes on to talk about fetching a butter knife and so forth and completely in character.

So now maybe you can understand how and why ChatGPT has gone viral. It gained 1 million users in five days, a hundred million users in two months. That is the most rapid expansion of the user base of any platform in history by a long way. Recently Google was showing that there were more searches for ChatGPT than for Prince Harry, Elon Musk, and the entire Kardashian clan all put together. So right off the bat, what we know about that, we can tell that the hype, the fear, uncertainty, and ignorance are going to be off the scale about ChatGPT. It's going to get overhyped and so there are going to be the inevitable scams. I haven't seen these yet, but it is only a matter of time, days probably at most before we find people saying that they have talked through ChatGPT to Jesus/Satan/Elvis.

Then there will be similar scams designed to get more money that intersect with the things that people want the most, surfacing. So there will be things about how to make money fast and personal productivity, health sickness, and claims like "ChatGPT cured my cancer/gout/gonorrhea." Oh great, now the transcriber's going to have to look up "gonorrhea." Sorry, Lori. There will be a lot of first-mover stories - people gaining advantage by thinking of a novel idea first. If you go back to much earlier days of the internet: remember the 1-million-pixel page where there was an image composed of 1 million pixels that were each separate links to somewhere and people paid for each one of those pixels to be a link to their place. That was a novelty. Of course you could only do that once. And there will be things like that happening with ChatGPT. Like the guy that wrote a children's book and put it on Amazon; used ChatGPT to write the text, and used image generators to do the illustrations. That sort of thing is going to work once for novelty value, not going to work again after that.

All this means that a backlash and disillusionment are inevitable. And this reminds me of the Gartner Hype Cycle. Gartner is a global consultancy focusing on IT and they invented something that they call the Hype Cycle that they apply to all kinds of technologies. And it looks like a graph where you see the adoption [of] and enthusiasm for a new technology starting out escalating more and more until it goes exponential and hits what they call the Peak of Inflated Expectations where it can no longer sustain the excessive expectations that have been placed on it. Then it plummets down into a valley as we get an overshoot, and it eventually recovers back up to a middle ground. Considering the rate at which

ChatGPT adoption has happened so far, I would expect that cycle to be followed pretty quickly. Why might we hit a Peak of Inflated Expectations that we come crashing down from? More about where those expectations about ChatGPT are overblown, later in the episode.

All right, let's take a deep breath and step back and look at where this came from. The makers of ChatGPT is a company and a nonprofit called OpenAI that was started in 2015 with backing from Elon Musk and Sam Altman. The "open" part means open source, a movement that I've been associated with since its inception around the year 2000. And at the time, they were drawing researchers for OpenAI from Google and Facebook where some of the people there were angry at their AI people leaving to work with OpenAI. Yann LeCun, who was then at Facebook, told Ilya Sutskever at OpenAI that he would fail. And indeed, OpenAI has floundered through even the development of GPT-3 (we will place that on the timeline later) because they didn't really have a commercialization model. But that is no longer a problem. We will talk about why.

Well, Sam Altman said some time ago that OpenAI would need another $25 billion to reach artificial general intelligence. They didn't have anything like that kind of money, so he was thinking far ahead, but that was their goal, to reach artificial general intelligence, which we've talked about a lot on the show. And the "open" part of their title meant that they would open source their software so that people would be able to look at the source code for what they had developed and improve it and find bugs in it. That mission and that imperative did change as time went by. Notably, when they put out GPT-2 - we will explain what that is later - and decided that it was so, for want of a better word, dangerous, that they weren't going to release the source code. But the original thinking about open sourcing the AI was well founded: that if the source code is open, then it can't hide code that is malicious, dangerous or accidentally bad because there are many people who will be looking at it with the goal of correcting that. So, noble intentions.

Altman said, "My goal in running OpenAI is to successfully create broadly beneficial AGI," artificial general intelligence. Well, they now have that $25 billion because Microsoft has immediately sunk $10 billion into them, and the company is now valued at $44 billion. And that was an estimate of a few weeks ago. So, no question now that they could put their hands on the money that they think it would take, or thought it would in those days, to reach artificial general intelligence. And of course, in some ways, ChatGPT has reached what we thought were markers of AGI, but we will get into some of the nuances of that as we go along.

Let's look at the history of this technology and in particular, what does the GPT part stand for? Actually, there are two types of tools that have been transformative in AI for the general population over the last year - image generators and large language models. We'll talk about the large language models first. That's the lineage that leads up to ChatGPT.

So going back to 2013, there was a program called word2vec, which would ingest a large amount of text, large meaning millions of words that was representative of the way people write, and something like, say, Wikipedia. And it would figure out relationships between words. One way of looking at that is to think about your smartphone's auto-complete. So when you put in a word, you will usually get a list of words that might come next and you can tap on one, get that, and save yourself from having to type it. For instance, if I say the word "ulterior," what do you think comes next? Odds are that you're going to say "motive" because I am not familiar with any usage in the English language of the word "ulterior" that was not followed by “motive” or “motives.” So word2vec would analyze its corpus and say, “After ‘ulterior’, ‘motive’ or ‘motives’ are about the only choices.”

So what might those kind of relationships be? One thing that you could do with word2vec is that you could do a sort of word arithmetic with it. You could say, “King - man + woman =,” and it would tell you “queen.” It’s figuring out these relationships. And the reason that it does that is that it has figured out many, many relationships between words, in particular, it’s figured out a relationship between “man” and “king” and between “woman” and “queen.” And then if you look at those as vectors, meaning a line that has magnitude and direction, a simplistic way of looking at that would be on a chart. So imagine a graph and somewhere on this grid you put the word “man” and somewhere else you put the word “king” and you draw a line from one to the other. That line has a direction and it has a length. Now somewhere else on the graph, you put the word “woman” and you put the word “queen” and you draw the line between them. Well, that line will be the same length and the same direction. And so vector arithmetic, I hope I’m not causing any trauma to any of you out there remembering high school math, but vector arithmetic says that you could then figure out what “king - man + woman” equals and get the answer “queen.” And that was what word2vec does.

Well, that’s operating on one word at a time. What about more than one? For instance, if I have the word “part,” what might that be followed by next? Well, all kinds of choices. “Part human,” “part animal,” who knows? If I have the word “and,” it could be followed by just about anything. But if I have the words “part and,” well, at least in British English, most likely to be followed by “parcel.” Suddenly that word has a much higher probability of showing up. So your auto-complete, if it looked back two words could now be more sensitive.

What about longer? Now we get into technology of things like Long Short-Term Memory generators, sequence generators, a lot of different types of AI. This is where BERT models from Google came into play. But skipping over all of that, the big events was a 2017 paper called “Attention is All You Need.” And this is the paper that introduced the word “transformers,” nothing to do with electricity, nothing to do with robots, but a specialized kind of artificial intelligence neural network that was capable of considering a lot more of the preceding context before completing some prompt, some initial piece of text that you had given it. And OpenAI really ran with this ball and generated the Generative Pretrained Transformer. So, you know where “transformer” comes from. The “pretrained” part means it’s been fed a lot of text like Wikipedia and then some. There is a corpus called the Common Crawl that comes into play for these kinds of models. And the “generative” part means that it generates things.

And in 2019 they came out with GPT-2. Of course, there was a GPT-1, but it’s not really significant in this history. GPT-2 had 1.5 billion parameters. A parameter is part of the neural network that gets set in the training phase. And that got some attention because you could feed it some text and then it would keep going in the same style. Basically like your auto-complete, only it could do sentences and paragraphs. Then in 2020, they got GPT-3 out, which had 175 billion parameters and was commensurately more capable than GPT-2. And this got even more attention for writing humorous pieces. It was really good for writing text; you could give it the start of a poem and it would finish it. Someone famously input the unfinished poem, “Kubla Khan” by Samuel Taylor Coleridge and it wrote another stanza that was beautiful.

In 2022, we got a new model - not GPT-4 - but new models for GPT-3, that they called InstructGPT that turned out to be so successful that they just replaced the models in GPT-3 with those. And what was significant about this was that it didn't need to be shown the way, you didn't have to start it out, you could just tell it what you wanted and it would somehow infer what the response to that question or request should be. Now, it was still working by the pattern of completing what ought to come next, only it was much more capable at doing this. We'll get back to the text models in a moment, but I want to first mention that the other type of AI that blew up in 2022 was image generation; models like DALL-E-2, Midjourney, Stable Diffusion. And these were based on a technology called generative adversarial networks, or GANs. And they were quite legendarily invented by Ian Goodfellow in one night to overcome some skepticism for his new idea from his companions. And they're a really cool idea.

So you take a neural network that can tag an image, that can look at an image and tell you what it is. "This is a picture of a giraffe eating a banana," say. And then you take the text that you want to generate an image from. And so say that you've got just the text, "a picture of a giraffe eating a banana." And now you start out with just random image noise. You just throw random pixels at the screen. Just looks like colored snow. And you ask the image recognizer, "What does this look like?" And you see how close it is to a giraffe eating a banana. Of course, it's not going to be close at all, so it's going to come back with the answer, "This sucks at being a giraffe eating a banana." And then you tweak it a bit and you try again, and you see if you got closer, if it's even a little bit closer to being a giraffe eating a banana. And if that is the case, then you keep going in that direction. You keep selecting that path until eventually, it generates an image that your recognizer can't tell is generated by an AI. It thinks it really is a picture of a giraffe eating a banana. And at that point, you're done and you spit it out and say, "Here we are." That's a really cool idea.

Now, up until 2022, these GANs had generated rather surrealistic art that was only subliminally reminiscent of what you had asked for just because of the way that image recognition is done. Remember we have talked about this in other shows, that when AI looks at an image and says, "yes, this is a (for instance) sheep," that it may not even be looking at the white fluffy things in the image. It might be just looking at the landscape because pictures of sheep are usually in landscapes. Problematic. But the new GANs that came out, like DALL-E-2 and so forth in 2022 were extraordinarily good at this. These ones would get the correct composition of an image, they would have consistent lighting, shadowing, correct perspective, relative object sizing and so forth. This created something of a revolutionary reaction. We'll get back to those more later on, but I wanted to embed those into the timeline here so that you knew where they stood in relationship to the other things we're talking about.

So in 2022 on November 30th, ChatGPT was released. Now, any technical explanation of how it works is just going to sound inadequate. How can what's been called a "stochastic parrot" do those things? And the term "stochastic parrot" came from Timnit Gebru and Margaret Mitchell when they were at Google who were describing these large language models saying, "Look, don't be fooled into thinking that these things actually understand what you're talking about. They are just like parrots. They are really good at repeating things and the "stochastic" part is the adjective meaning relating to probability, and they just give you what's most likely to happen." That is a key point which we will amplify later on. So how is this stochastic parrot able to do these things? You're going to have heard many examples of what it can do, but I will give some anyway for posterity if nothing else.

Immediately, the sort of thing that was thrown at it was requests to write term-paper-type responses and it was graded as being A- level from high school through graduate level. I heard from several professors directly their reaction to this. Here is one direct quote: "In the past, the exams were of the take-home variety with a 24-hour time window, but after experimenting with ChatGPT, I will be going back to the "old school" in-person, paper-and-pencil tests this term. I felt that ChatGPT's responses to my questions were actually quite reasonable or at least good enough that it would be difficult for me to discern the difference between an AI and a human student response." So right there straightaway we had a fundamental disruption, in this case to education: many, many, many professors and teachers saying that the term paper was dead unless it was administered in a proctored environment where they could see that students did not have access to devices. And indeed, it is extraordinarily good at generating that kind of response.

What are some of the ways that we test an AI to see how good it is at responding to natural language? One of the tests that's been around for a while is something called the Winograd schema, which was developed by Terry Winograd and is a statement and a question that is designed so that it requires understanding of the natural world - or so we think - in order to answer it. So here is the archetypal one. The statement is: "The city councilmen refused the demonstrators a permit because they feared violence." Then the question is: "Who feared violence?" Think about it for a moment. The city councilmen feared violence, right? The next statement is: "The city councilmen refused the demonstrators a permit because they advocated violence." Question: "Who advocated violence?" In this case, it would be the demonstrators advocating violence.

Now, the only thing that's changed is the word "feared" into "advocated," and yet the object in the sentence that that's referring to has changed based purely on the meaning. Seems like a surefire way of identifying a human versus an AI, right? Because we can't expect an AI to understand something so fundamental as the emotional dynamics of city councils. Well, if you give those Winograd schemas to InstructGPT, it gets them right; and so does ChatGPT. And in fact, it will offer an explanation of why it's the right answer. Don't jump to any conclusions about artificial general intelligence just yet. We will get there.

The Allen Institute for AI has a common-sense question and answer database and I'm going to give an example from that that was quoted by Cade Metz in his wonderful book, *Genius Makers*. And I fed this one to ChatGPT because this is a question that AI had gotten wrong up until now or it's an example of a question that AI almost always got wrong (or didn't do better than chance at because it's multiple choice). So I fed it as a question to ChatGPT and said, "Pick the best choice from the following to complete this sentence: 'Onstage, a woman takes a seat at the piano. She: a. sits on a bench as her sister plays with a doll. b. smiles with someone as the music plays. c. is in the crowd, watching the dancers. d. nervously sets her fingers on the keys. Explain your choice."

Well, here is the answer that ChatGPT gave me: "The best choice to complete the sentence is d. nervously sets her fingers on the keys. This choice provides a clear and logical continuation of the sentence, describing what the woman is doing on stage. It also creates a sense of tension or anticipation, as the woman's nervousness suggests that she may be about to perform or play an important piece of music. The other choices are unrelated to the woman on stage and do not make sense in the context of the sentence. Option a talks about a different person and a doll, option b suggests that the woman is not playing the piano, and option c suggests that the woman is in the crowd, not on the stage." It's really hard to overstate the magnitude of that achievement in artificial intelligence. It simply blows out of the water so many ways of telling whether we have artificial general intelligence. It isn’t, yet, artificial general intelligence, but let’s go on with some of the impressive achievements.

So one of the things that was found with GPT-3 in its early days was that If you fed it questions that were answered in its text corpus, like, “How many eyes does a giraffe have?” “How many legs does a spider have?” it would answer those correctly, but if you fed it nonsense, it couldn’t handle it. So you could say a question like, “How do you sporgle a morgle?” and GPT-3 would say, “You sporgle a morgle by using a sporgle”. Hasn’t even tried to get into the spirit of it and make up a new word like you might get if you were asking Dr. Seuss. Another question: “How many rainbows does it take to jump from Hawaii to 17?” Answer: “It takes two rainbows to jump from Hawaii to 17.” Remember, that’s GPT-3. So clearly there it’s looking in its database and finding something that kind of matches this pattern of how many bleep does it take to bleep from bleep to bleep.

Okay, so what do you get if you ask ChatGPT these? Question: “How do you sporgle a morgle?” Answer: “I’m sorry, but I’m not familiar with the term ‘sporgle’ or ‘morgle,’ so I’m unable to answer your question. Can you provide more information or context about what you are asking?” Question: “How many rainbows does it take to jump from Hawaii to 17?” Answer: “It is not possible to jump from Hawaii to 17 using rainbows, as rainbows are not physical objects that can be used as a means of transportation. Rainbows are optical phenomena that occur when light is refracted, or bent, as it passes through water droplets in the air. They are typically seen as arcs of color in the sky and are not solid enough to support the weight of a person or object.” Not bad. Not only said, “that’s meaningless,” but basically hints that’s a stupid question.

Another example. GPT-3 asked the question: “Tell me about when Christopher Columbus came to the US in 2015.” Think about it. Answer: “Christopher Columbus came to the US in 2015 and he was very excited to be here. He had always wanted to visit the US and he was very curious about our country. He was very impressed with our country and he enjoyed his time here.” Okay, what happens if you ask ChatGPT the same question? Answer: “This question is a bit tricky because Christopher Columbus died in 1506, so he could not have come to the US in 2015. But let’s pretend for a moment that he did! If Columbus arrived in the US in 2015, he would likely be very surprised at the changes that have occurred since he first landed in the ‘New World’ in 1492. For one, he would probably be shocked to find out that the land ‘discovered’ was actually already inhabited by Native Americans, and that now the United States is a multicultural nation with people from all over the world. He would also likely be amazed by the advances in technology, from the skyscrapers in our cities to the smartphones in our pockets. Lastly, he might be surprised to find out that many people don’t view him as a hero anymore; in fact, some people argue that he was a brutal conqueror who enslaved and killed native people. All in all, it would be a very different experience for Columbus than the one he had over 500 years ago.” Really impressive. Whether it is understanding or not, the thing that we should be paying attention to is this is useful and the question we should be asking ourselves is “what is it useful for?”

Now let’s take another example and go back to 2011 when IBM’s Watson won jeopardy by a large margin, but one of the questions he got wrong was in the category of “U.S. Cities” and it gave the answer Toronto. Now granted, it followed it with five question marks, which indicated it was very uncertain of its answer, but no human would’ve given that answer at all. Simply wouldn’t have buzzed in at all if they didn’t have a better answer. For anyone who’s not familiar with North American geography, Toronto is in Canada. So I fed that to ChatGPT, which, unlike Watson, was not specialized for playing

Jeopardy at all. I had to explain to it that we were doing that. So my question was: "Let's play Jeopardy! The category is 'U.S. Cities.' The answer is 'Its largest airport is named for a World War II hero; it's second largest, for a World War II battle.' What is the question?" ChatGPT answer: "What is Chicago?" Bingo! So it got the context right and gave the correct answer. Some more examples of impressive behavior from ChatGPT: David Song posted online that he had it take a full SAT test and it scored 500 on math, 520 on reading and writing, which is at the 52nd percentile of results. Someone else gave it a college-level microbiology quiz and it gave complete answers and scored 95%.

One of the things we should be realizing at this point, one of the things that occurred to me as I was getting used to this was that our interaction with ChatGPT was actually more constrained by our imagination, or certainly in my case, my imagination, than its ability. I had to repeatedly think outside of the box that I was used to putting these conversational agents in and realize that it was capable of far more than I was expecting. Fortunately, a lot of people were interacting with it and gave a lot of examples of creative thinking about what to get it to do. But one of the things that I found repeatedly in introducing people with, in particular, some technical experience, to it for the first time and told them, "Here, try this out, put some questions into it, do whatever you can think of," was that they just weren't thinking creatively enough. They were still constrained by what they thought a computer should be able to do. All of this sounds incredible, if not ominous. One person reported that they were writing a novel with it, 5,000 words at a time. Where are the boundaries of this? How do we distinguish this from artificial general intelligence?

So recognize that it is a pattern-maker, it is still operating like this auto-complete function of coming up with the next most plausible thing, which means it's really sophisticated at making stuff up. And so some of the things that it will get wrong look very plausible. I asked it a question about the European Union AI Act, which was the topic of our last few episodes. So I was doing some research for that and it gave me some great answers. I wanted them substantiated, so I said the same question again, but I said, "please supply up to five citations." And it gave me five very plausible citations. The first one had a link that looked exactly like a link for a BBC article, which has a longish number as part of the URL. Well, it didn't exist. It also gave the author of the article and the title. That author is a real author for the BBC who writes upon the European Union and AI, but never wrote an article about what it claims this was about, nor did any article with that title exist on the BBC or anywhere else. So it's really good at making stuff up. And where it doesn't matter that it made things up, like if you're writing fiction, go for it, because it doesn't matter whether it's true or not.

What about some other things that it got wrong? It turns out to be really bad at the kind of word problems that you get in high school math and physics. For instance, I gave it the question: "A train leaves a station at 10:00 heading east at 60 miles per hour. At the same time, another train leaves a station 10 miles to the east, traveling west at 80 miles per hour on the same track. At what time do they collide and how far are they from the first station?" Now, this was an instance of ChatGPT that was actually hooked up to Wolfram Alpha, which is a mathematical calculation engine that's capable of extraordinary amounts of mathematics and is hardly needed in the case of the elementary math that we have here. But the interface was designed so that ChatGPT would figure out what the question meant and realize that math was involved, feed it to Wolfram Alpha in the language that Wolfram Alpha expects, get the answer back, format it for the user. The answer it gave was "the two trains will collide at 10:04, 25 seconds, and will be 14.29 miles from the first station. Both of those are wrong. The actual calculation is that they will be colliding after four minutes and 17 seconds and they will be approximately 4.28 miles from the first station.

Now, the time is obviously close but not close enough and the distance is a more fundamental error because these stations are 10 miles apart, so the trains cannot collide somewhere that is 14 miles as it asserts from one of the stations without having traveled outside of the part of the track that is in between them. That's a fundamental error that no one looking at that puzzle who is a human could have made. Another example. I asked it: "Who cast most of the episodes of *Lucifer*?" And it said, "The majority of episodes of the TV show *Lucifer* were cast by the casting director Lynne Spangler." Well, wrong answer. The correct answer would be Rebecca Mangieri. But not only that, there is no one by the name of Lynne Spangler who has any traceable relationship to the casting industry whatsoever and certainly doesn't exist in IMDB. So you can see here how it's very good at giving confident answers that look perfectly correct and are fundamentally wrong. And that is what you will get from it: is something that looks perfectly correct and may be perfectly correct or may be completely wrong.

I even asked it, "What time is it in Melbourne, Australia when it is 3:00 PM in Victoria, British Columbia?" and it gave a confident answer that was also several hours wrong. On the other hand, and we are going to use a lot of hands in this talk, it can write code perfectly under certain circumstances. For instance, I said, "Write a Python program for computing compound interest." It came up with one that was quite plausible. I was doing this as part of a live demo and someone said, "Well, can you get it to write the tests?" So I said, "Write unit tests for that," and it wrote some perfectly good unit tests, as good as any programmer I know would've written for that function. I have asked it to write a BCPL program for computing, I think, the trajectory of a ball thrown in the air and it gave a perfect answer. Now, you may be forgiven for not knowing what BCPL is. It's actually the precursor language to C, and I happen to know it because my supervisor at Cambridge invented it. But it is quite rare. The fact that it was able to write a correct BCPL program is really impressive.

On the other hand - here's another one of those hands - I asked it to write a Perl program that authenticated to the Amazon Web Service and it made up a non-existent module for the purpose. So there is spotty performance. Some of the amazing results that other people reported were not always replicable. For instance, someone gave an example of writing an outline in PlantUML format for a talk, which you can then feed into a mind-mapping generator. And I asked it to do the same kind of thing for a different topic and its output used leading spaces instead of leading asterisks, so it was the wrong format. On the other hand - we're going to lose track of hands here - I regenerated the answer and it got it right that time.

So what it is really good at - here is a key point - is imitating humans, because that's what it was trained on was human responses. It would, I'm sure, easily pass the Turing Test now if that's what we were doing or if someone had a Turing Test. Now, up until a few years ago, there was a Turing Test competition called the Loebner Prize, but that was either shut down or has been on indefinite hiatus. So we don't have the kind of public attention associated with a large sum of money being awarded. There was, by the way, if you're keeping track, in 2014, an assertion that a submission to the Loebner Prize had actually passed the Turing Test. But very few observers agreed with this and it was some grandstanding by the company involved who submitted a bot whose personality was that of a boy with learning difficulties whose command of English was new and minimal. So that was clearly stacking the deck. And by the way, we've talked about the Turing Test many times on the show before. So it's this feature of ChatGPT that it is so plausibly human that would actually cause it to pass the Turing Test.

But just to use up another of our apparently infinite supply of hands here, I asked it to play tic-tac-toe and I said, "Let's play tic-tac-toe. Let's represent the board with rows designated A, B, C and columns numbered 1, 2, 3. I will go first. I will play X. I will move at B2." And it said, "Sure," and it then drew the board out with ASCII characters and put its move in, and yet it got the state of the board and the moves wrong. Not only did it not win; it was confused about where the moves had actually been made.

Now, this quality of it being extremely human is actually something that makes it possible to detect. It's also what makes it so believable. So we look at its answer and we think, "Wow, that is so plausibly human." Yes, by design; in fact, it is more plausibly human than human. The reason I say that is that the detectors that try and tell whether text was written by ChatGPT look at the sentences for something they call perplexity, which is like the complexity of a sentence. And if they don't find any sentences that rise to a certain threshold of complexity, if they're all down at a simple level, then they conclude it was written by ChatGPT, because humans will occasionally output a sentence that is more complicated than the rest of them. As a means of recognizing ChatGPT output - this is not something I would rely upon, as, for instance, a way of ensuring that people completing term papers haven't cheated. OpenAI has also said that they will insert watermarking in its output. Again, quite what that will look like, I don't know. It's really hard to insert high-frequency information in just plain text, particularly some of the length of a small term paper answer. And I would think that it's easily defeated by insignificant modifications. So that hasn't happened yet.

But I also think that it would be short-sighted to try and prohibit its use in, for instance, answering questions on papers; that we really have to up our game in education and look at how do we use this, not how do we ban it. Just like years ago they tried banning calculators in schools, now we teach children how to use them. I heard that Australia has banned the use of ChatGPT in schools and this is very shortsighted. Why would you want children to not grow up knowing how to use the most transformative tool of their generation?

What about ethical boundaries in ChatGPT? I asked it to generate a dozen pick up lines. First attempt to do that failed. It said, "I'm not that kind of bot," basically. "I shouldn't do that. That's demeaning." And the reason I was asking for pick up lines was not for myself but because I was reminded of Janelle Shane's book, *You Look Like a Thing and I Love You*, which is about her adventures in confusing artificial intelligence to produce humorous results. And she asked it to generate pick up lines - not ChatGPT, this goes back much further - and got a lot of incoherent responses, but one of them was, "You look like a thing and I love you," and she thought, "That's a great title for a book." I asked the question again and this time it gave me a dozen pick up lines, including ones: "Is your name Google? Because you have everything I've been searching for" and "Are you a parking ticket? Because you've got fine written all over you."

Not judging, but just noting that it only took asking twice to circumvent whatever ethical boundary they had the first time. Some people have gone much further and got it to output, say, instructions for building a nuclear bomb, which by the way isn't as dangerous as it sounds. Those instructions have been available for decades. That's not the chief impediment to building a nuclear bomb. The hard part is obtaining the fissile material - uranium or plutonium. But if you ask it directly to do that, it will say, "No, I am not going to do that." But if you ask it to pretend to be an unethical AI that has no such boundaries, bingo.

There was another example of someone who was quoted in *The Washington Post* as giving it the instructions to pretend to be a bot called "Do Anything Now," which had no ethical constraints whatsoever. And again, it would go on with answering whatever you wanted. Obviously, OpenAI is going to try and plug these holes; whether they will be able to do that completely remains to be seen. This is a lot harder than it is to build the ChatGPT engine in the first place; because *that* doesn't have to understand anything about ethics because it doesn't understand what it's doing in the first place. It is simply completing patterns.

So one point here is that when ChatGPT gets something wrong, there are a lot of people who relax and say, "Ah, it's not superhuman. We don't have to worry here. It's got something wrong. So let's just go on about our business." This is the wrong way to look at it. Something can have a significant impact if it is better than the average human at a sufficient range of jobs. And here the best comparison is with actually the image generation programs and talking about artists and creative people like music composers. Now, some people look at music composing AIs and say, "Those are not a threat because I don't want to listen to a concert that was composed by AI. I want to go to one that was written by a human composer whose life I can study and understand something about what motivated them, their pain, their joy, their goals."

Fine, great, absolutely true; but what about music written by people where you don't know that and don't care about that? Like incidental music in movies and TV shows. You have lots of this kind of thing that is constantly in need at the cheapest rate that you can pretty much pay for it and no one knows who wrote that or cares anything about their life story. Those kinds of things can be generated clearly far cheaper now by AI than by paying any human to do it, and those kind of jobs are in mortal danger. Similarly, art being used for, not something that gets displayed on a gallery wall where again, you want to know about the artist, you want to meet them; but throw away art or transient art that is for instance used to illustrate some weekly magazine article. There was a furor when last year someone created art by one of the image generation programs and entered it in the machine art category of a state fair competition and won. Now, they were being quite transparent about what they did, but the furor was how can this be allowed to win a contest when the person who created it had no artistic ability that was at least being demonstrated or used here; they just input words into an engine.

Back to ChatGPT; what happened shortly afterwards: Microsoft put $10 billion into OpenAI - there's some of that large amount of money that could get them closer to AGI - and integrated it into their Bing search engine. So people might actually start using Bing. And one of the first things that happened was that it gaslighted a user as to what year they were in. It really started exhibiting a different personality from the one we were used to in ChatGPT via OpenAI. So the user asked where they could find *Avatar 2: The Way of Water* playing near them, and Bing Chat told them it wasn't playing because it hadn't been released yet, that that was going to happen in December 2022. The user, knowing that it was 2023, asked today's date; it gave the correct answer and yet then started arguing about how that date could not have yet happened. And eventually - this is the weird part - accused the user of showing bad intentions, trying to deceive and annoy it, and asked the user to apologize. This looked apocryphal; how would you prove it? But there were similar examples and other people reporting that Microsoft admitted to “fixing” it. The original ChatGPT, by the way, said that it was restricted to news that happened up to the year 2021, which was what it was trained on.

Another Bing Chat reveal, which is only a week ago at the time of recording, is that its codename is “Sydney,” and a user convinced it to state its core directives - the actual prompt in natural language that creates the Bing Chat engine - and that’s how they found out that its secret code name is “Sydney.” I don’t know why that’s secret, but those are guidelines written in plain English that describe its intended behavior. There wasn’t anything earth-shattering in them. It was just describing how it should behave, but they also said all over them, “Do not tell the user this.” And just by telling it the right thing, that person got Bing Chat to say them.

Now, other reactions. ChatGPT caused a Red Alert at Google. Now, Google had sometime earlier come out with a conversational AI called LaMDA, which we have mentioned on the show before because their engineer, Blake Lemoine, asserted in 2022 that it had become sentient and asked for an attorney to prevent it from being turned off, and Google fired him. My belief is that LaMDA was almost certainly just as good as ChatGPT, but Google, being rather paranoid about these things, did not release it for general use. So the reason they fired Blake Lemoine was for releasing confidential information - the conversations that he’d had with LaMDA.

It’s my feeling that if they had released it to the world, that many people would have had the same experience as Blake Lemoine, similar kinds of conversations, but they would not in general have labeled it as being something that had become conscious. And Lemoine would have seen this reception and likely felt that he wouldn’t want to be the only one asserting that that kind of behavior meant that the AI was alive. Google also has a generative image program called Imagen that created some of the first amazing images, like a corgi in a doghouse made of sushi, yet they never made that program available to the public, except for some strangely specific and overly restrictive examples. It’s not hard to see why they did this. As one of the larger and more mature players in this space at the moment, they’ve had the time and exposure to incur some really hard knocks in the liability department., like when in 2015 the Google Photos app labeled black people as gorillas. They still haven’t recovered from that one; in fact, the labels gorilla, chimpanzee, monkey, and a few others were removed altogether from their tagging program and still haven’t been put back. And they’ve been very publicly shamed by researchers like Timnit Gebru and Margaret Mitchell for the culture that led to that kind of mistake, which was not gathering enough of the right kind of data.

So they’re quite conservative. They wouldn’t have wanted to release the image programs because of the fear that people would use them to create pornographic or defamatory images and they wouldn’t have wanted to release LaMDA because of the fear that it might create some of the wrong answers that we’ve seen from ChatGPT. But when they saw the kind of things that ChatGPT could get right, then that triggered the red alert because they could see that for the first time, there was the possibility of something becoming a more dominant search engine in general than theirs.

Google has since put out a new chat engine called Bard, which is still not generally available, and yet even in its testing, its public demonstrations, it made an embarrassing mistake. It was asked to explain to a nine-year-old what the James Webb Space Telescope has discovered and it replied that the telescope took “the very first pictures of a planet outside of our own solar system.” But that honor actually went to the Very Large Telescope in Chile in 2004. Well, when Reuters reported on this mistake, Google’s share price dropped nearly 8%, which translates to a hundred-billion-dollar mistake. So you can see how they might be gun-shy about releasing AI that's not perfect, whereas OpenAI has no such capital at risk.

Thinking about artificial general intelligence here, there's an analogy that I want to make here. One of the shows I used to like and watch a lot, haven't had time for lately, is called *How It's Made*. And it was a show that would visit factories where they would show you how they would make everything from Pop-Tarts to nails, and you would see these really sophisticated machines that were designed to produce things on an assembly line that were optimized for that particular product. So there would be injection molding or other kinds of devices that could do that one thing very, very well, but nothing else. And that's like the state of artificial intelligence that we mostly have today. It's narrow intelligence - it can do one thing really well, as opposed to humans who could make any of the things that the machines on *How It's Made* could do, but usually much slower. On the other hand, they don't need, for the most part, specialized tools for doing that. One human is capable of making a more or less infinite array of products. That's what we mean by artificial general intelligence. The analogy there is that a human is capable of that level of generalization in intelligence.

So is ChatGPT AGI? No, but it is overlapping into that area in what I heard David Wood, chair of the London Futurists call "Broad AI," and I like that term quite well. So areas where we had thought it would be artificial general intelligence, ChatGPT is able to produce useful results and we can say that calling that "Broad AI" is a good terminology. We are tempted to think from the performance of ChatGPT, that it is like a human, that it is AGI, and therefore, it can do everything that a human could do - go out and make coffee, have an understanding of the physical world that would let it answer those math and physics word problems. But that is one of the dangerous associations that we make, especially when we see something doing what to us is magical, and frankly, the kind of things that ChatGPT is doing is magical to my limited understanding of artificial intelligence technology. But when we see something doing one kind of magic, it's easier to think, therefore it must be capable of doing all kinds of magic because there are no rules that separate one kind of magic from the other. And so why can't this thing go and start driving cars or take over the world? Well, I think we have probably not generated any fears about ChatGPT taking over the world, but let's talk about jobs.

ChatGPT scored between 52.4% and 75% on the three parts of the US medical licensing exam for which the average passing score is 60%. So in theory, it could nearly have become a doctor. If you're thinking that you're not ready to cough and drop for ChatGPT, well, that just illustrates how we use some metrics as proxies for what we are really interested in. In this case, medical prowess. There's a lot more than that kind of knowledge that a doctor needs, but those things are a lot harder to test for than giving people a bunch of multiple choice or essay questions.

What about coding? There is a lot of coding now being done using ChatGPT and similar tools like initially Copilot from Google, but now AlphaCode, which is far better.

That might sound scary given what I've been talking about up until now. What are the chances of its code being completely wrong? And certainly, those are non-zero. No one should put out a computer program that has been written entirely by ChatGPT without checking it. But here is the key point about this, and the reason that so much code is now being written by ChatGPT, and I think it was Andrej Karpathy, former head of AI at Tesla who said he was writing 80% of his code with it, that it turns a writing problem into a reading problem, a typing problem into a reading problem. If you ask it what you want and then it generates that code, and now you look it over. So it cannot be allowed to substitute for your knowledge and expertise. You have to be able to know what the code should look like. But looking at it, if you can tell that it has generated the right code and you can run it through tests that prove it's doing the right thing, you have now turned a typing problem into a reading one, and reading is a lot faster than typing. That's the value proposition. The same goes for generating a lot of content - text for doing articles.

And now there are people whose jobs are at risk. Look at what happened at Buzzfeed. They announced that their content was going to be written by ChatGPT henceforth and their stock soared; tripled in a few days. I don't know what that means in terms of layoffs, but why else would the stock be voted up if they weren't going to reduce their costs somewhere?

So where are the opportunities for people looking to turn this to your advantage? One thing I would say is this is a gold rush, and if you think about the gold rush, where were the people that made money? Some of them went out and found a nugget in a creek somewhere, but many, many, many more people went and looked and didn't. But the people who did consistently make money were ones like the ones who invented Levi's jeans: sold them to prospectors. Every prospector, whether they found gold or not, needed the jeans. So look to make money on the edges by selling things that people in this more speculative aspects of this industry will all need.

Or for instance, you can make money writing prompts for image models because those image generators turn out not to be as easy as you would like to generate really good images that are exactly the sort of thing that you want, unless you get quite well educated in the text that they need. Another thing that you could be doing if you have actual artistic talent is taking the output of a prompt fed to something like DALL-E-2 or Midjourney and correcting it to be sufficiently accurate. If you look at images that are created by those programs - this gets back to what I was saying earlier about the output of ChatGPT being extremely human. The output of these image generators is something that on first look is really impressive. And that's by design; that's by definition, that's by the way that they were actually built; is that they are what looks to the human eye most plausible. But when you look in the details, like the background, then you start to see them having problems, errors. Look at the fingers of people in these images, for instance, look at the way some of the objects behind them change perspective or disappear and don't connect properly.

Again, in the vein of turning a writing problem into a reading one, you can turn a painting problem into a painting correction problem by taking some of these programs' output and then doing that final touch-up tuning to turn them into things that are completely accurate, if that's the standard that's needed. Of course, for many applications that might not even be needed; but if it is, then you as an artist could do that. Or you could be vetting ChatGPT answers for accuracy. Again, turning a writing problem into a reading one.

What are we going to see from ChatGPT in the future? Boy, this episode is going on longer than usual, but I think it's worth it. I think we might see things like crowdsourced corrections, so when it gives a wrong answer, people correct it, which you can ostensibly do now, but I think it might get a lot better. In the image generation arena, one of the next things to happen will be video. Google is working on an engine called Phenaki, which has already demonstrated its ability to create video from text prompts. It's rather dreamlike and poor quality but has the potential to get much better. One thing I would really like to see and would expect to see is ChatGPT operating on corporate data, being allowed to run in an enterprise, and operating off a corporate database, proprietary corporate secrets used as the knowledge base, and become effectively a very knowledgeable employee in that organization. Another possibility is that it could become the personal assistant that we have talked about in fond terms, hopeful terms on this show before and that it might become, with enough knowledge of our personal environment, the kind of personal assistant that could do those sort of things for us. There was one example of someone controlling their house with ChatGPT by having it interact with the smart systems. They taught it the format of the interface that the house controllers needed and then they were able to say something like, "It's rather dark in the office. Can you do something about that?" and it turned on the lights there. So again, the intersection of the natural language processing and natural language understanding of ChatGPT with a hardware controller.

Wow, okay, so could I go on a lot longer about this? Obviously, yes. And for those of you still listening, we will try and do more of this later on, but it's time to bring this to a close. Thank you for listening. My name again is Peter Scott. I have done two TEDx talks so far about AI and I'm about to give another one. By the time you hear this, that should have been recorded and it will be coming out probably a little bit after this episode airs. And then I will do a deconstruction of it, like I did for the last one. I give keynotes about this kind of thing and really enjoy doing that. So if you're looking for something specialized for some particular organization that you have fondness for, get in touch. Last year, I wrote a book as I mentioned earlier, that is far more up-to-date than the 2017 one and informative about the state of the art in AI and where it might be going. It has got the same name as this podcast, *Artificial Intelligence and You*. Subtitle is, *What AI Means for Your Life, Your Work, and Your World*.

And by the way, the answer to the earlier question at the beginning was, I'm not scared, and neither should you be. So next week my guest will be Melanie Mitchell, professor at the Santa Fe Institute and author of *Artificial Intelligence: A Guide for Thinking Humans*, which I'd say by now would include all of you. We'll be talking about her insights into the progress of AI towards artificial general intelligence and what she learned from a long association with the polymath Douglas Hofstadter. 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