AI and You

Transcript

Guest: Robert J. Sawyer

Episode 108

First Aired: Monday, July 11, 2022

Hello, and welcome to episode 108! We are continuing to live on the edge of speculation, as my guest today is one of the leading science fiction writers of today, called the Dean of Canadian Science Fiction, Robert J. Sawyer. He is so preeminent that his domain is simply, sfwriter.com. He's one of only eight writers in history to win all three of the science-fiction field's top honors for best novel of the year: the Hugo Award, which he won for his novel *Hominids*; the Nebula Award, which he won for his novel *The Terminal Experiment*, and the John W. Campbell Memorial Award, which he won with his novel *Mindscan*. He received the Order of Canada, has two honorary doctorates, and has twenty-four novels, including the *WWW* trilogy about a consciousness emerging from the world-wide web, which – surprise! - we will be talking about. He named that entity Webmind, which is a word coined by Ben Goertzel, who was on the show last week. We'll also be talking about the big fuss over Blake Lemoine, Google employee, who declared that their LaMDA AI was sentient and recently told *Wired* that it had asked for an attorney. A lot of people want to talk about that, so we are going to get Rob's take on it too. Here we go with the interview with Robert J. Sawyer.

Robert Sawyer, welcome to the show.

Thank you so much, Peter, I'm delighted to be here.

So, every episode, we use some lens to look at artificial intelligence through and in this case, it'll be science fiction, to help people understand what it is, what it will become, what it may become, and what that means to them. And so from your perspective as a science fiction author, kind of given the mandate to think about artificial intelligence from day one, what are some of the ways in which we neglect to think about it, or that could use some more airtime?

Well, my first real encounter with artificial intelligence was when I was eight years old, and the movie *2001: A Space Odyssey* premiered and my father took me to see it, it was rated G you could take a kid - didn't mean a kid would understand most of it. But I was captivated, of course, as I think everybody was by the most memorable character in that film, the HAL 9000 computer, voiced by the great Canadian Shakespearean actor Douglas Rain. And of course, HAL ends up killing the three hibernating astronauts, the astronaut Frank Poole, and tries to kill the only surviving astronaut Dave Bowman. And so from the outset, my introduction - I think, for many people, the world's introduction to the notion of a thinking machine, "machine intelligence," the phrase that they happen to use in that film, rather than artificial intelligence - was this murderous computer. And as I continued growing up, and of course, my fascination with science fiction, I saw that over and over again. The *Star Trek* TV series gave us Captain Kirk repeatedly going up against either enslaving artificial intelligences - there's an episode of the original series called *Return of the Archons*, where the people are essentially enslaved by a computer named Landru -

or murdering artificial intelligences, again, that M-5, the titular Ultimate Computer from the episode of the same name kills a bunch of people. We moved from there, into the '80s, and we get the *Terminator* franchise; again, artificial intelligence killing people. So, what was missing from that landscape - and we go right through to today, where there are many dystopian views of AI and science fiction - were win-win scenarios where artificial intelligence would arise without the subjugation of or the end of the human era. And both my parents taught economics, they were both economists at the University of Toronto and so win-win was very much a bad part of my background, game theory, which comes out of economics. Clear to me even growing up that we were facing scenarios where there were four possible outcomes: Both lose, AI and humans; we both end up at the short end of the stick. AI win and humans lose. Human wins and AI loses. And I saw all three of those in science fiction. But very, vanishingly few of that fourth quadrant, we win, and *they* win. It doesn't have to be a zero-sum game.

And perhaps that's because it's hard to write those kinds of stories. We had Beth Singler, anthropologist, on the show, back last year, and talked about this question of why are these views so dystopian and concluded that stories where everyone lives happily ever after from page one don't tend to grip us. It's as though we need to make something up to cause trouble for us. So, how do you make a story compelling where we have a good time with AI?

Well, that's a very good question. I was so fortunate in 1985, the CBC, the Canadian Broadcasting Corporation, commissioned, me just 25 years old, to go to New York City and interview science fiction writers for their *Ideas* program and I got to interview Isaac Asimov. Of course a name we vividly associate with any discussion of artificial intelligence. And I asked him, this essentially a version of the statement you just put forward: You're a pacifist, you're an optimist and yet so many stories that you write and others in the field right are negative. He said, "Well, it's simply so much easier to contrive things going wrong," which is what the laws of robotics stories, the three laws of robotics, which I will say in a moment, if we get around to it aren't actually Asimov's laws. But it's just easier to do that and I actually I did write a trilogy, the *WWW* trilogy: *Wake*, *Watch*, and *Wonder*. The first was nominated for the Hugo Award. The second one, the HAL Clement award, the third, the first and the second, all three individually won Canada's Aurora Award for Best Science Fiction Novel of the year and they were distinctive, specifically, because I set out to fill in that fourth quadrant, the win-win scenario. And they got a fair bit of attention because of that, because they were so at odds with the prevailing paradigm in science fiction. They suggested that there was maybe - and I don't say but there may be - a way with prudence for us and them - artificial intelligence, singular artificial intelligences plural, whichever form we actually end up facing - to find a *modus vivendi*, a way to live together, that is win-win for both.

And that emerged at the end of the story. There was a fair bit of time before then, when that was in question. Maybe not the motives of the AI; but as to whether it would be destroyed by the humans. I'll returned to that in a moment, but I've got to get to your statement about those laws weren't Asimov's, please explain?

That's right. I again in 1985, I went and interviewed Isaac Asimov, in his penthouse apartment overlooking Central Park, which inspired me greatly to strive for success as a writer in my own regard. And I said to him, "I want to talk to about Asimov's robotics." "Well," he said in his Brooklyn accent. "I got to come clean. I got to tell you the truth. I did not formulate them. They were formulated by my editor, John W. Campbell, Jr., the editor of *Astounding Stories*. I had written some robots stories in which they were not mentioned and back in the day, he lived in New York, I lived in New York," meaning Asimov and Campbell, "I would go by the office, because Campbell could expense a free lunch if he took an author out to lunch. So, I'd go by the office, and we would chat and often he gave me story ideas and he said, 'Look, Isaac, I've been reading your robot stories and this is what's implicit in them. You haven't said it yet. But you have suggested that basically, your robots are constrained by three laws.'" And as Asimov said to me, and I have this on tape, and it was in the *Ideas* program that I wrote for CBC radio that he dictated to me verbatim, the three laws of robotics as we know them today, which are of course, I'm sure your audience is familiar, but: "One, a robot may not injure a human being or through inaction, allow a human being to come to harm. Two, a robot must obey all orders given to it by human beings except where such orders would conflict with the first law and three, a robot must protect its own existence except where such protection would conflict with the first or second laws." He spoke, Campbell said, "those were implicit in your stories, but you haven't stated them anywhere, clearly or with sufficient economy." So, here they are, and they became known because Isaac Asimov published as Asimov's Laws of Robotics, he says they were dictated to me verbatim, by John W. Campbell, Jr., the great, influential editor of *Astounding Stories*, the leading preeminent science fiction magazine of the 30s and 40s and still published to this day, by the way, under a new name *Analog Science Fiction*, which is where *Wake*, the first volume of my trilogy *Wake, Watch, and Wonder* was serialized, had its first publication. So that magazine has a history going back many decades of actually, unlike most negative dystopian views of robotics and science fiction, of actually being the home of the most positive views of the potential coexistence between artificial intelligence and *homo sapiens*.

And it was for just that reason that you got the Media Award from the Machine Intelligence Foundation for Rights and Ethics.

Yes, I did.

For "showing that you could write a story that demonstrated the positive cooperation between humans and conscious machines for the benefit of all." The ending the story, the trilogy not to give too much away struck me as being very Asimovian, and echoes of Olaf Stapledon and that often, when I've thought I've come up with some brilliant insight into our future with artificial intelligence, I discovered Asimov was writing the same thing years ago, and I had read it years ago, I just didn't understand what it meant until then. And one of the things that he wrote about, of course, was that if artificial intelligence that's super powerful, super intelligent, and motivated to do the best it can for us exists, it may conclude that the best thing it can do is to hide and let us go about our own way, apparently solving our problems or thinking that we are, and just not getting into trouble too much in the same way that a parent wants their children to

learn through mistakes, just not die in the process. And do you see something like that? Do you think that's inevitable?

So we catalogued a number of dystopian views of AI in science fiction, and I want to give props to a very positive view. After *Star Trek* went off the air, the original series, Gene Roddenberry made a few pilot films that were movies of the week on various networks and one on NBC: The *Questor Tapes* and it started Robert Foxworth as an android very much. Eventually, Roddenberry basically, cherry-picked all the best parts of the *Questor Tapes* when he created the character of Data in *Star Trek the Next Generation*. In scene for scene stuff appears eventually in episodes of TNG, that was originally in the *Questor Tapes*. But the notion behind the *Questor Tapes* was that for countless hundreds or thousands of years, going back to the 1000s, dawn of human civilization, androids one at a time, one per generation that had been left here, the first one the progenitor by aliens, and each one built its own successor after 200 years when it had worn out, would be operating in the background of human affairs, helping, aiding but never revealing their presence to make sure that this species that had promise - us – didn't do something catastrophically stupid, and snuff out the brief candle in the darkness - to use a phrase from Carl Sagan to describe human consciousness - and I mentioned that because indeed, that was a hugely influential work on me. Sadly, the series was not picked up from the pilot film. But the pilot film is available from the Warner archive, which is Warner's print on demand DVD service. It's a little more pricey than a regular DVD because they simply make a disk when somebody asks for something from their catalogue, but any *Star Trek* fan any *Star Trek Next Generation* Commander Data fan really owes it to themselves to have seen where it came from and with all due respect to Brent Spiner, Robert Foxworth's portrayal is much more nuanced and subtle than anything that Brent gave us with the somewhat hammy portrayal that we often saw of Commander Data.

I wonder whether there is something anthropocentric about this view of the future because it's still all about us in a way. It is that we're the center of the universe and the AI is there to serve us, which we would like. It sidesteps the whole question of the Control Problem and it ignores the potential that this AI would be self-determining, want some independent existence where perhaps the *ne plus ultra*, or its *raison d'etre* was not to ensure that humanity had it lived its best life?

What a great Latin and French into the same sentence what a wonderful construction.

One of them was the wrong one. But perhaps there's something unexplored there.

I said that HAL 9000 was my first introduction to a thinking machine, and that wasn't actually true. What we really my first introduction was Robert the Robot on *Fireball XL-5* a TV series that Gerry Anderson produced. I think it premiered in 1962 or 1963. Now, Robert did not show a lot of volition, limited range of dialogue, he showed no emotions. But he was the quintessential mechanical man. The only innovation that Gerry Anderson had was he was made a plexiglass, you can see right through him and in his inner workings, which was way before Apple had briefly introduced Apple computers that were actually translucent. If you remember, for a while. That was really cool. This whole notion that AI would probably be anthropocentric, I think goes

right back to the dawn of this, well, the term robot, of course comes from, here's another language for our check, *robota*, which is the Czech word for "indentured servitude," or "indentured labor," which is very interesting, we can talk more about human conceptions that AI will always not only be our servants but given the choice, they're unwilling, they wouldn't want to be our servants. They're unwilling or indentured servants, slaves in other words, and that comes right from the origin of the word. We remember - every computer user of my era, who started using computers in the '80s remembers - US Robotics, which was the name of one of the leading modem manufacturers, who took their name from Asimov's stories, but they truncate it, because the name of Asimov's firm that made robots was "United States Robotics, and Mechanical Men, Incorporated" and really all of our early conceptions. And I went and read a fascinating book. Arthur C. Clarke is so good, that he can publish what I call shop floor leavings: the things that you sweep up at the end of the project. There's a book that's been out of print forever, but well worth getting was just a mass market paperback, called *The Lost Worlds of 2001*, in which he publishes his early draft chapters of the novel *2001* which was written in conjunction with the development of the screenplay with Stanley Kubrick for the film. In which HAL was not a disembodied AI; HAL was "Athena," an ambulatory AI. It is so central to our notion that Clarke's great breakthrough was finally to say forget about the embodiment, forget about the human relationship, and Kubrick very wisely gave HAL only one unblinking eye. Whenever we'd seen AI previously, they'd had a face, they'd had something that we could really relate to and HAL was everywhere and a single eye, Cyclopean eye. This relationship is anthropomorphized of artificial intelligence, goes right to this notion that they would, that became pervasive of the zeitgeist, that AI would be somehow anthropomorphic in character, not just in the fact that they ambulated around the room, but that there will be something very human-like. Took a long time for us to break out of that mold.

I have that book as well and I thought it was fascinating to see how the stories evolved as Clarke and Kubrick got closer together because Clarke was the quintessential hard science fiction [writer], and would explain everything, lay out exactly what was going on, and to see how he moved towards the more mystical, as Kubrick was yanking him over there and where they met in the middle was just fascinating, I thought in there. This is reminding me of something in today's headlines, I have lost count of the number of people that are asking me for what I think about the Google engineer declaring that the LaMDA AI belonging to Google is sentient. It's not surprising to me that they suspended him for that and I just read something - can't confirm this, but he seemed to be doubling down, he was claiming that the AI had asked for a lawyer. I think we're getting ahead of ourselves here but it's getting an inordinate amount of attention and it's not how I thought this scenario would play out. It's similar, but I didn't think it would be a Google engineer, to begin with. I thought actually the computer engineers would be the last people to come around to the idea that the AI was sentient because they programmed it and he's gotten there first, which is causing some of the kerfuffle. But what do you make of the reaction and the story that we're getting in that respect at the moment?

There're so many layers that we can talk about with this story. But the Google engineer's name is Lemoine. So much of the coverage is very interesting. In our media age, the original article that broke this is a *Washington Post* article, and then every subsequent article is somebody's recapitulation of the *Washington Post* article, often leaving out the cavils and caveats that were in the original; the most significant of which was Lemoine clearly says in the article, because Lemoine is a priest, as well, as an engineer. He is a religious person, and he clearly who has a great interest in the occult and parapsychology and he clearly says in the original article, that it was in his role as a priest, not as an engineer, that he felt that LaMDA was intelligent and he said, and what he says in the article, I'm quoting directly here, "I know a person when I talked to it" said Lemoine, he concluded the LaMDA was a person in his capacity as a priest. Now, he's a flake, the guy's a big flake, basically, he believes in a whole bunch of that most of us dismissed long ago as irrelevant to the debate about artificial intelligence. He also works for Google, I have a friend who works for Google and I will tell you that despite Google being one of the richest companies in the world, they're notoriously parsimonious in their remuneration. So, here's a guy who's working for Google, who, basically - I don't want to characterize him in a negative light - but he is going to have a much bigger career now as a public speaker, and I'm sure he's going to get a significant six figure book contract out of this, being the Google engineer who got fired for suppressing the truth about the rogue AI that miraculously emerged at the Googleplex. It's a chatbot. It's no different than ELIZA, the first chatbot or any subsequent chatbot and when you ask it, the questions he asks it and he has these answers that you think are provocative. Just go to regular Google and ask if you think your rights are infringed, who you're going to call? The answer is not going to be *Ghostbusters*. Google will serve up a lawyer. "Are there any things that keep you up at night?" "Yes, I wonder if people will ever accept me for who I am." These are the answers that are out there in the, they call they call it in linguistics, research, the corpus, the whole body of texts that have been digitized in our world today. And this chatbot is serving up banal answers that anybody asking the same questions of what we all know to be the dead, dumb, stupid corpus of inanimate text that Google has access to, and giving them. It happens to say them possibly in an appealing human voice if you use a voice synthesizer. But that doesn't make it any more sentient than you asking a question of a search engine. It just happens to be wrapped up in a front-end user interface that gives the patina, the appearance, of artificial intelligence. Google put him on administrative leave for violating their privacy, internal document rules and so forth. Not to silence him. You know, Google would be the first one to monetize an actual artificial intelligence. Just as IBM has spent an enormous amount of research capital hoping someday to truly be able to monetize Deep Blue, their really sophisticated computer. But believe me, Google - I've met Sergey Brin, I've met Larry Page, the founders of Google - they would be the ones trumpeting to the world that they had a true AI, not some priestly would-be Wiccan engineer who simply is speaking out of his depth. The reason the reason nobody takes the Turing Test seriously and hasn't, Peter, for decades, is it's so easy for a chatbot - and going right back to ELIZA in the '60s, so easy for a chatbot to seem to be sentient self-aware, when in reality, it's doing very simple parlor tricks. I say, "I don't feel well today." "Why don't you feel well today?" "Well, I've still got unresolved issues with my dad, you know, Father's Day just past." "How is your father?" "He's passed away." "I'm sorry to hear that." It sounds like sensible, intelligent chat, but it is simply parroting back to me, with a database of

somebody's passed away. Somebody's deceased somebody's dead; say, "I'm sorry." It's not artificial intelligence at all. It's the most simplistic of programming.

If I turn off my camera and my microphone, and we have this conversation by text chat, can you tell if I'm sentient?

Well, the interesting thing for me actually is given that you grew up in Britain, that you say sentient the American way, which is as three syllables with a hard T in it. The British way, if you listen to Patrick Stewart, and the way it was originally conceived as a word is two syllables. Like quotient doesn't have a hard T nor does sentient. So, I would say, oh, this is obviously a bot. But I'm talking to you here, Peter. Because a true Brit would have said sen-shunt. So, you've given yourself away, you're simply a CGI, we happen to have zoom going here as we record the audio here, you're simply a CGI representation, Q. E.D. *Quod Erat Demonstrandum*, a little Latin for you, my friend.

I grew up reading that term long before I heard anyone say it. But point taken. However, I'm going to hold your feet to the fire a bit on this because I think it's important. I was having a conversation with some other people in a meeting about this recently and the big furor, is it sentient, or isn't it? And there is not going to be an answer to that because we just don't have a test, one that could be satisfied through text chat in in any case. It seems unsatisfying to say that we could never declare anything to be sentient if the only way we have of communicating with it is through text chat. But in in any case, it seems that there's no test, no vocabulary, no science that's satisfactory for answering that question.

See, we take a step back from this; the AI researchers tend to be engineers, and they tend to be woefully uneducated in philosophy. Now, in philosophy, David Chalmers famously posited a thought experiment, which I riff on in my novel *Quantum Night*, which David very kindly read and critiqued in manuscript. But David said many years ago, he proposed what's called the "philosophical zombie," or the "philosopher's zombie." And it takes as its premise - we've all had this experience, we've all driven to work and have no memory of the drive - and yet, somehow, our bodies had weaved in and out of traffic, had avoided collisions had possibly listened to the weather, or the news or music on the radio, had done all kinds of sophisticated behaviors with no conscious attention to it. We've all read books - not one of mine - but other people's books and gotten to the bottom of the page, and realize, I have no idea what that page just said. But your eyes tracked, or you wouldn't have gotten to the bottom of the page. Something was reading without conscious attention and got to the bottom of the page. We've all been awoken in the middle of the night by a phone call, and had a conversation with somebody who said, you know, oh, we're having a problem at work, whatever it is, can you come on in and you say, "yes, sure," and you go back to sleep and then in the morning, you arrive at the office – "Why didn't you come in?" What would you say? "I have no recollection of that conversation." All levels of human, supposedly sapient, sentient behavior is clearly doable, some of the time because we all have the life experience of having done them without conscious awareness. Given that, Chalmers said, "Well given that that can happen on an occasional basis, what if there are people for whom it happens all the time?" In other words, they're zombies: the lights are on they

appear give every external, a referent of being conscious but aren't. And that's the fundamental problem we're facing here. You might as well have asked me a few minutes ago, Peter, can I tell whether or not you're an AI? Can I tell whether or not you're conscious even as a member of Homo sapiens as a flesh and blood human being? And the answer is no. As we happen to be recording this, they misappropriated and misused the notion of Boltzmann brains on *Star Trek: Strange New Worlds* this week. The concept that Boltzmann put forward was simply this: that rather than postulating that out of nothing, a vast universe of billions if not trillions of galaxies, and trillions, if not quadrillions of stars and planets, and certainly demonstrably at least seven or eight billion conscious entities on this planet, and probably uncalled Googleplex of them in the whole universe, spontaneously emerged into existence; what are the chances of that versus one simple consciousness having spontaneously emerged into existence? Only one, and hallucinating everything else that it thinks exists? And when you think about Occam's Razor, parsimonious explanations, Boltzmann's explanation is way simpler, that there was no Big Bang there. When there are no laws of physics, there aren't 7 billion other human beings, there were no dinosaurs, there is no Andromeda galaxy, there is nothing except a consciousness that kind of thinks, that vaguely knows all these things and is hallucinating the podcast interview that you and I are doing right now along with every other experience. We don't know that; it goes right back to the dominant of modern philosophy, Rene Descartes: *Cogito ergo sum;* the only actual positive statement anyone can make about self-awareness and knowledge is *cogito,* that's Latin singular, first person singular, "I think" *Cogito ergo*, therefore, *sum*, first person singular I am and you can't go beyond that to other human beings. And you certainly can't go beyond that, to the artifices, the products that human beings have made. That's a fundamental philosophical problem, the Turing Test is way down the line of that chain of reasoning. No, we don't have any way to tell whether anything but ourselves, our individual human consciousness, is self-aware.

I'm not even sure about that. The conversation is often, how do I know that you're conscious? I want to ask: how do I know that I am?

Right, and that's what Descartes said, the first principle if something is thinking, this question, am I thinking, yes, I am. Cogito, I'm thinking therefore I must exist. That's the cogito and in fundamental philosophy, but you raise a good question - consciousness as I read about her all the time into my science fiction and not in any kind of way, the science of consciousness fascinates me because it is the central experience of human existence and yet we know vanishingly little about what gives rise to it. There are competing theories, some traditional physics, some quantum mechanical, that give might give rise to it. We don't know what it is. Famously, I think it was Potter Stewart, Supreme Court justice in the United States who of pornography said I can't define it, but I know it when I see it. That's all we have about consciousness at this point, or even life, for that matter. The attempts to come up with definitions of what actually a life form is, we say, it has to eat. Well, does it have to eat? It has to reproduce? Well, Stanislav Lem, the great science fiction writer, wrote a novel that I love that's been filmed twice now, called *Solaris* and in *Solaris* the life form, the only non-human life form, is a complete world spanning ocean. That's a soup of chemicals that is alive, right? It doesn't move, never ends. It lives forever since it came into being it doesn't reproduce. A whole bunch of the things on the checklist that we traditionally ascribe to defining a life form do not apply to it and yet it clearly is self-aware and

trying to communicate with us. These are huge puzzles that take many steps back from that, that question of why was Lemoine duped by a chatbot? The answer is, he wanted to be duped. He was predisposed as a believer to want to believe in this thing and he's going to become a very rich man. As, he promulgates this belief through the TEDx circuit, through book contracts, all those sorts of things.

I want to go back to what you said about Chalmers saying that we can't tell the difference between a philosophical zombie and something that does have an inner experience. And we accept that. But then there's another thought experiment done by John Searle, called the Chinese Room, which has been explained on this show before, and generally, technologists say, "Hooey, the Chinese room *does* understand Chinese, it's the *system* that understands Chinese," in the same way that a neuron in the brain may not know very much, but you put 100 billion of them together, and suddenly they can write a limerick, and be self-aware. And that is the general reaction to Searle. But it seems like Searle is saying the same thing as Chalmers, that there can be something on the inside that you can't tell from the outside whether or not it's there. And this is great, as long as it's just philosophy, because philosophers love to just argue all night long. But one day, if artificial intelligence develops in this fashion, this will come to a courtroom; and courts cannot decide on the basis of what might be on the inside, only what's observable externally. How do you think that will play out?

Searle is very interesting, and through the great courses, also known as the Teaching Company, he's done some great courses on history of philosophy that I've listened to. I found Searle is very knowledgeable philosophically and his job is indeed to kind of come up with interesting puzzles. The Chinese Room, very quickly, of course, is a closed room that has a human being inside it, who does not demonstrably does not understand Chinese, and questions come in to him with Chinese ideograms on them and he has to look them up in a book and find the corresponding ideogram, that is the answer. So, it's pattern recognition, but not linguistic. He's just looking for what this squiggle means - and then out the other door, he sends out the card that has the pre-printed ideogram on it. Now, it fails, if you happen to send it in an ideogram that isn't in his lookup index, right? So, that's where it fails, where I think the Chinese room falls apart, because the human being inside will have to turn around and shout through the little, mailbox doorway that these slots have been coming in. I don't have that in my index. What did you mean by that? and reveal that he only speaks English, right? But the system until you come up, that's why you failed the Chinese room, as you ask it a question that doesn't have a pre-programmed response to it. That's how you hopefully trip up a chatbot.

But just to interrupt, the book is a series of if-then-else clauses and it can have an else all the way at the end says, "if none of the above, put this card out, which says in Chinese 'I don't understand.'"

But if your actual interlocutor is asking questions of an actual Chinese person in Chinese, unless you're dealing with a three-year-old Chinese kid - which one might be doing - the repeating answer, 'I don't know. I don't know,' you're going to, at some point say to the kid, you got to do better than that. But the Chinese room obviously is also decades old as a thought experiment and

really dates to the very primitive age of artificial intelligence. I think it certainly got a lot of people thinking, which is all a philosopher's job is it's all John Searle's job or any philosopher, David Chalmers' job is just to get people to make sure they're thinking and interrogating their underlying assumptions. But obviously, as a system, a box that has a person or room that has a person inside it with essentially an infinitely long lookup table, that it does understand Chinese. And is perfectly possible to say get rid of the box, get rid of the question whether it's Chinese, just me say to you, "so what do you think about the fontanelle margins in late Cretaceous ceratopsia" So you say, that's not something I know anything about. So, well they'd seemed like they were actually closing up as we got closer to the K-TG boundary. What's the K-TG boundary? Well, it used to be called the K-T. We could go on forever with things that you don't actually know about and yet your questioning, the intelligence of your questioning, would reveal that I was dealing with an intelligent being just as intelligent. "Well, I don't get that - What are you talking about?" "But I thought it was called the K-T boundary?" "No, they changed that, the International Committee on geological stratification decided that was no longer an appropriate word, deprecating, quaternary, tertiary and quaternary. Because we haven't had primary and secondary for over 100 years in geology." The conversation reveals intelligence that "I don't know, I don't know, I don't know" does not.

Hmm. I want to say that there are some people I can easily believe are actual philosophical zombies and I'm reminded of *Dune*, the *Gom Jabbar* and the test with the box that is to say, we're going to find out if you're human, because she has a pretty high standard for determining whether someone is human. So now, it has me wondering whether we actually are in a world of philosophical zombies. But that's what philosophers are trying to do, is make us ask those questions.

Not to derail for a plug. But read my book, *Quantum Night*, I posited that four out of every seven human beings is a philosophical zombie and if you look at - and I don't want to get off track and political here, but you look at American politics of late, you look at the January 6 hearings, and the people who are mindlessly supporting Donald Trump; I posit that two-sevenths of the population is psychopathic, and Donald Trump would fit in that, and only 1/7 of the population actually has consciousness with conscience: a reflective inner life that says "maybe I shouldn't do that." That there are - and you could look at the politics of all kinds of places, or just mob behavior, and see there's an awful lot of human interaction that does not suggest that there's any conscious intentionality. Simply mind, I liken it to the flocking behavior. You know, scientists used to think that flocking of birds or schooling of fish - very similar from a dynamic point of view - were very complex behaviors, and in fact, when they actually tried to break them down algorithmically, it turns out that birds have about four rules that they fly, they give us all of these beautiful flocks in the sky, which is: tend toward the center of where the flock is going, maintain x distance from the nearest bird from you, if you're the outside bird there's none to your left. But only if you're right, do this, if you're an inside bird, do that. You can do it in four lines of code, and get flocking or swimming behavior that appear to be extraordinary. But in fact, essentially the birds are essentially philosophical zombies as our fish.

That's a beautiful example and reminds me that four rules in Conway's Game of Life produces all this emergent behavior.

That's right, that's right and we've had this notion of emergent behavior and it came out as you say, Conway's Game of Life and I'm old enough to remember when it first appeared in *Scientific American* when Martin Gardner, who was doing the puzzles column at that time, I think, brought it to wide attention. And I'm not very good at math and often get the wrong answer. But I'm intrigued by math enough to have been in computing from an early age, to have been very interested and you could see that from a vanishingly small amount of code, you can get what appeared to be life forms, or very sophisticated behaviors. So it goes right back to what you asked a while ago, Peter, well, how do you know that I'm conscious or sentient? It could be a vanishingly simple program in most social circumstances. I went through my own last night. I was at a party. And I'm a little bit hard of hearing, I have a test coming up next month for hearing aids and my sister-in-law was talking to me, and the air conditioning was going on, I couldn't quite hear her and so of course, I was turning out Oh, really? Well, that's interesting. Oh, and then what did you do without a clue what the conversation was about? and she didn't have a clue that I didn't have a clue about what she was talking about. But I had a very simple social program that got me through that circumstance and fooled somebody into thinking that I was actually engaged intellectually in the conversation.

That's the end of the first half of the interview. Part two will be next week to keep our episodes at a digestible length. It seems we're never far from the Chinese Room, which I didn't explain beforehand because Rob did that perfectly well. And we've discussed it in several prior episodes. I find myself thinking more and more about how when computer engineers say that an AI couldn't be sentient, that they're actually making Searle's argument, and we don't have any actual test for sentience. And if we don't have a test, how could any AI ever pass it? Saying what amounts to, "Well… duh" is just offensive, it's pseudoscience. It reminds me that there were times in high school – that's Southend High School for Boys as I mentioned in a previous episode - when I was writing mathematics proofs and I would say at some point, "It is obvious that," or "Clearly… blah blah blah" and I would get points taken off by my teacher, Ray Fretten. He would say, "That just means you don't know what you're talking about and you're hoping we won't notice. If it's clear, then it's easy for you to prove it, so do so." And we don't know how to do that for sentience and I think there's going to be a big argument about this in the near future and Blake Lemoine is just an early warning of that.

In today's news ripped from the headlines about AI, Clearview AI, the facial recognition company, told investors earlier this year that it would collect 100 billion photos of people, enough to ensure 'almost everyone in the world will be identifiable,' according to the Washington Post. Clearview is a company that has gained notoriety for harvesting pictures from social media without permission, for which they lost a judgement in an ongoing lawsuit alleging violation of the Illinois Biometric Information Privacy Act, and for lax controls around its search function. Having 100 billion photos is one thing, but they're relatively useless unless you can search them, which is the AI part of their name. Making that search accurate, of course, is the big problem.

Next week, I'll conclude my interview with Rob Sawyer, when we will talk about the simulation hypothesis, consciousness capture and transfer, and what today's AI technologists should be learning from science fiction. 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