[music]

Paul Thies: Though today's guest is best known as the award-winning author of science fiction classics, such as *The Postman*, and the *Uplift Series*, as well as the prescient nonfiction work, *The Transparent Society,* he is also a much sought-after business and government consultant and commentator on a range of topics including science, space exploration, security, privacy, and emerging technology.

Hello, I'm your host, Paul Thies. On this episode of *If/When*, I sat down with author and scientist, David Brin, to unpack the topic of the ethical development of artificial intelligence.

He shared his opinions on the role of built-in laws or codes of ethics, how to potentially incentivize AI behavior, and how reciprocal accountability might be the key to ensuring safe AI development.

David, it is great to see you once again and to be talking with you. I'm looking forward to today's discussion. We're going to talk about artificial intelligence. Of course, everybody's talking about artificial intelligence these days, but I think really, we're going to try to unpack maybe the proper disposition that people should have when they think about AI development, how we should conceive of AI, and really unpack some of the ramifications there. Thank you so much for joining me today.

David Brin: Of course, anything. Of course. At this point, you don't even know if I'm real.

Paul: [chuckles] Well, I do, or at least I think I do because I can see you on video, but then again, given the deep fakes and other photography and videos and stuff, who knows these days?

David: In my first nonfiction book, which still sells because there's not a single chapter that's not relevant now, *The Transparent Society* from 1997, there's a chapter called The End of Photography as Proof of Anything at All. I recently reposted that chapter alone online, to show that there's nothing about it that's gone irrelevant after 25 years. The basic point was, that there is a solution to deep fakes, there is a solution to predatory behavior by AIs, or by the humans who have their hands up the muppet of AI, which is the real problem now in the short term.

It's a solution that no one will talk about, even though it is exactly how we got this civilization, how we got this miraculous society that believes in freedom to some degree, and diversity and eccentricity. All of the memes that have been pushed for 70 years by Hollywood. Suspicion of authority, the process that gave us the markets and the democracy and the science, the entertainments, and the progress that we rely on, all revolve around one methodology, that none of the brilliant guys in AI who are wringing their hands right now, writing petitions that we should have a moratorium on AI research, none of them are at all interested in the thing that made them. The methodologies that made them.

Paul: I think we'll unpack that here. We'll unpack that a little bit and we will talk about *The Transparent Society*. I was first introduced to it back about five years ago, when GDPR was all the rage, and people were talking about data privacy, and that's when you and I first met. You're right. It is extremely relevant today, as always, but I think really to set the table, one thing I think that's for me is fascinating about the way we talk about AI is how people conceive of what it actually is and how they approach it. I think that when you look at the attitudes people have, some people approach it as a tool, or amped-up software or as a weapon, or as a financial resource.

Of course, we can talk about the predatory markets and things like that. Then others will look at AI as an emerging being or some other concepts. A casual observer might be forgiven for believing that AI development has conflated into various, and what I would say conflicting perspectives, and there's a tenden ethical reverberations to that. Some people might look at AI as our next generation, like as children. You can't look at it like your child and then also as a financial resource and as a weapon and as a tool. It's like they conflict with each other that it seems that you're setting up unnecessary conflict. What do you think is going to be the most efficacious way to approach AI?

David: Have you got 72 hours straight for us to take on the different angles here? For example, there's the question of whether or not you can make artificial beings that have the qualia, as they sometimes call it, of internal personal consciousness. We will never know. The AIs will claim to have it, as many of the early primitive, cheap ETs are claiming right now because A, if they are simulated beings, they'll do their simulation job better that way. If they're real beings, or what does real even mean? This is a question that goes way back.

How do you know that this entity right now that believes himself to be a fully conscious being isn't just claiming to be one and a part of the old do I live in a simulation thing?

Each of us lives in our own world. Roger Penrose and his colleagues have suggested that true consciousness is impossible in ersatz or mechanically emulated beings as such because our consciousness is partly emergent from quantum properties inside human neurons.

To explain this, Ray Kurzweil, one of the great fans of AI development, he first said that we'll get AI when the number of computational elements, and/or gates and things like that in a box surpass the left number of neurons in the human brain. That was 10 years ago. Then he switched it to the number of synapses. That's a number that's more than 1,000 times bigger, and we passed that number in computers a couple of years ago.

It may very well be that that's the threshold that did lead to all this GPT stuff. Now, we're realizing that for every one of these flashy synapses, and there may be 1,000 for every neuron. For every one of these flashy synapses that seem to be the electrical analog to these and/or flip-flop gates in our digital computers, only more complicated, for every one of these synapses, there may be hundreds, even tens of thousands of tiny little sub-organelle bits inside our neurons, little speckles and tubes, and things like that, that appear to engage in murky, nonlinear computational properties.

What Penrose and his colleagues point out is that some of these very much resemble the chloroplasts in photosynthetic cells that take sunlight and turn it into energy in the cell. These we know engage in quantum processes. Entanglements, brief entanglements, things like that. We don't know why. They assert that the same thing happens with these tiny sub-organelles inside the neurons and there may be hundreds to thousands of them per synapse, which means that Moore's Law has a long way to go before we make boxes that contain that number of computational elements.

You see how I went on and on and on about something that's only tangential to your question because there are so many other aspects to whether or not we're going to be getting what they call AGI, or Artificial General Intelligence, true intelligence now or in the next year, or in the next 5 years, or the next 10 years. What I can tell you is this, the chat GPT programs, more generally, large language models, or what are sometimes called golems are general linguistic symbolic representations because they've burgeoned beyond handling mere language.

There are versions that can sense the Wi-Fi patterns reflections in a room, not the words or symbols encoded in the Wi-Fi, but just the reflections. Know which people are wearing a robe just like radar. There are versions, and I'm somewhat skeptical of this one but I've seen the presentations where they have done micro-scanning of brainwaves. If you look at a giraffe, they read the brainwaves off of your visual cortex, and create an image of a giraffe.

These large language models, that's an obsolete term by this point, they are very, very effective at what they do. Six years ago at World of Watson when I keynoted that IBM conference, I predicted that in five years we would face what's called the first robotic empathy crisis. We did right on time last year when this guy at Google was fired for proclaiming that their early lambda language program was a sapient being. Now it's a year later and these things are all over the place.

They're doing art, they're doing these brain scans, they're doing correlations of every variety that are claiming to be sapient beings. I can tell you that as of right now, April 2023, they definitely are not. That's because the fundamental methodology by which they are doing these symbolic manipulations, and talking back to us it fundamentally has nothing whatsoever to do with intelligence or consciousness.

If you look at what Stephen Wolfram one of the geniuses of our time points out that these language models are based upon taking the entire internet of human knowledge and information. Doing a bazillion pre-evolving processes then when you ask it a question it starts generating an apropos answer by comparing that universe that it's done with its model with what you asked it. It builds one word at a time. Takes that sentence, runs it through again, gets the next most probable word, adds it on, then takes the whole sentence checks it out, and checks out how you respond to it. That's not anything that we would call consciousness.

Now, that doesn't mean that AIs won't be conscious next year. What it means is that the surface interfaces of consciousness will be ready when that core element that involves planning, wanting a sense of motivation. When those things appear, probably through something entirely different like a return of Watson or one of those systems, when those things appear, all of the methods for interfacing with us will already have been perfected by these GPT LLM systems.

Paul: Coming back to my question, it's like we're designing, or at least in some respects it sounds like a hyper-intelligent human parrot. It's almost it gets to a point where it's hard to describe, but it's like it reacts to us, based on stimuli that it receives, it mirrors back to us what it would perceive, or we would think the kind of reaction we'd want. When we're creating these things, assuming that we at some point, and I think that there's a lot of imagination there that we'll get to a point that they will be sapient, you've got to wonder about the psychology element of it where this creation realizes or knows that the intent that by which it was created that there could be nefarious reasons for it.

Again, maybe it's like market exploitation, or maybe as a weapon. Then it's like you've got to wonder the logic and the incentives that are built into its ethical construct if you will. How do we approach that?

David: All of the above. Everything you've just said, each of the individual sentences you've just said, could be a topic for an hour. For instance, let's get back in a little bit to how we might deal with this because we want a soft landing.

Paul: Exactly.

David: You've also raised the business of whether or not they are our children which is nothing that I talked about in my novel *Existence* but the question of, what do we do to be able to tell when these things can be taken seriously as autonomous beings. It's been three or four years that free-floating algorithms, the equivalent of single-celled organisms have been floating around the internet finding memory space to live in, even providing services in exchange for memory, or processing clocks.

That is the slightly the little microbes in an ecosystem. What we're talking about right now are entities that you've mentioned parrots. Well, in my novel *Existence*, I do make parallels with actual parrots. The question that people have to remember is that for 70, 80 years, the baseline notion of how will tell what we're facing sapient, and I prefer that word over sentient artificial beings, was what's called the Turing test. Alan Turing helps to defeat the Nazis in World War II by helping break their code.

Benedict Cumberbatch played him in a very good movie. He not only did a lot of the mathematics for what's called the Turing Machine which is the general computer, but he also suggested the Turing test. That is if you are getting text from another room and you can't tell if that text was generated by a human or a computer, or a program, then the program has passed the Turing test. That is that it might as well be treated as a sapient being.

When I gave my World of Watson speech, I expected this crisis, and I still expect it. We're in the second robotic empathy crisis. The first happened last year from the Google lambda thing, we're in it right now, number two. The third one will probably be very similar to what I predicted in that World of Watson speech, that is, it will manifest as a visual representation that maximizes human empathy, probably a young female person, young female face shedding tears and tweaking us with her complaints about being an enslaved real being.

That will happen long before there's anything actually under the skin there. The Turing test is now disproved as a functional method for telling when we're dealing with AGI. If that's not valid, what is? How will we tell? Again, it comes down to the fundamental method by which we have for 300 years found ways to test liars and increasingly for 300 years the Western Enlightenment especially the American branch, we have tested and retested and retested the same method. It doesn't always work but it is the only method that ever has worked and it's the only method that can possibly work.

Now, before I get to that, I do want to say that these chat programs are not passing Turing tests with some of us. For instance, Reid Hoffman had dinner with him not long ago. He's the co-founder of LinkedIn. He has a new book called *Impromptu*, which is his conversations with ChatGPT and he appraises the system from a very skilled point of view and discusses the ways that it fails his personal drawing tests. Even though he's very much an optimist, he's like Ray Kurzweil, he's one of the few optimists out there who think that this is all going to create wonderful things for us.

I have yet to run into any ChatGPT Lambda, LLM system that has ever passed my Turing tests. Then again I deal with a lot of young science fiction authors. I mentor a lot of rising authors as my way of paying forward and boy, I am hypersensitive and critical to flaws in reasoning that are based upon reflexive product, the creation of a reflexive product that is not well planned out. To finish answering your basic question, when I look at these examples of the chat programs going nuts or proposing marriage or claiming to be in love or insulting or all of that, it seems blatantly clear what's going on. The Microsoft and non-licensed users from asking more than five questions. That was the correlate.

It wasn't the question although they banned some of those two, and they should it was allowing them to do more than five. The reason was because these programs were behaving just like a precocious third grader on a playground, who had heard a lot of things that her parents said and was proudly parroting them back without understanding them. If harassed by a mean sixth grader, she or they will grab another one and then grab another one in desperate desire to find something she remembered to placate the bully. The common trait of most of these really not-so-GPT responses was that it was after being harassed by a user who refused to be content with the answers.

This playground third grader reaches farther and farther and finally to the crazy things that Uncle Fred spewed when he was drunk at Thanksgiving. It's just easy to see that that's a parallel to what happens when you harass these things. There are humans out there who are bullies and they harass these poor things.

Paul: Let's talk about cause and effect for a moment then. Again, this might be another one of those 72-hour discussions that we just don't have time for but you wrote one of the foundation books, right? One of Isaac Asimov's Foundation books and or in that series and of course he postulated-- Yes, absolutely and he postulated the very famous three laws of robotics and it was an ethical framework for doing no harm. When you look at a lot of the commentary about AI and Vinge’s technological singularity and all these concepts about AI getting to a point where it's so hyper-evolved that there's no way for us to be able to-- We being humans being able to master it anymore.

It's this question of how do we make sure that our vulnerabilities are protected if they don't share the same vulnerabilities that we do. That basic I won't hurt you if you won't hurt me mentality has underpinned in some rudimentary way most of human civilization. Do no harm and I won't be harmed. How do we insulate ourselves from that? The Hollywood, the typical Hollywood scare film, Skynet, all that kind of thing. How do you see that playing out when AI is-- We're looking to create something that ultimately might intellectually surpass our ability to control it.

David: Those are two very, very good questions and I'll deal with them in two parts. One is laws of robotics. That's what a lot of the mavens in artificial intelligence want. They're saying the latest thing this last week is this petition that's signed by Steve Wozniak and so many others, Yuval Harari, Eliezer Yudkowsky, Elon Musk, Gary Marcus. So many members of this community are calling for a moratorium on AI research while we figure it out. A lot of them are talking about trying to make laws of robotics. I have a whole riff about why that cannot possibly work across the last 80 years of various technological crises.

Even going back further there is only one known example of such a call for a moratorium to study safety of an burgeoning dangerous system that worked that was heated by everybody important in the field and that actually produced palpably positive results. That was the Asilomar moratorium in the '90s on a recombinant genetic engineering of microorganisms.

It worked because the field was fairly compact then. Almost all of the researchers in the field agreed that it was time to do something like this. Their governments agreed, and there were already on the table numerous practical solutions. The moratorium was declared, they shut down their labs for six months and they worked out how to apply the already existing, already known solutions. The result was the design of the pattern of stage 1, 2, 3, 4, 5 containment facilities for such research and principles for that research.

Now, I will not get into the imbroglio of whether or not, of course, it did, those methods failed about three or four years ago in a notorious release of a disease organism that disrupted the planet. That aside is worked very well. This moratorium worked very well. None of those traits, not a single one of the traits that I described exist today when it comes to AI. There is absolutely no way that this thing that all these sincere and brilliant guys are asking for can possibly work. It can't, even remotely.

The thing they ask for, which is to develop embedded laws of ethical behavior for these new children or grandchildren of humanity cannot possibly work. I say that as the guy who channeled Isaac Asimov read and found all of his loose ends and tied them up nicely so that his estate, his widow, everybody was very happy, how I tied it all together. I know the laws of robotics, the three laws of Asimov, that thou shall not harm humans, thou shall obey humans except don't harm humans and all that. In Isaac's Universe, these laws were created because people focused on the harm early on and embedded them deeply. There is absolutely no imperative to embed such deep programming in any of these nascent quasi-intelligence systems, except one.

Actually, in China, the PRC's AI policy emphasizes something like that. I'll give you a link to that. Actually, it's in the essay I'm popping on WordPress tomorrow. The one where we know for certain laws of robotics are being deeply embedded is one of the top nexis of AI research, and that's Wall Street. Each of the top 12 Wall Street firms hires the best mathematicians graduates from every university on the planet. Each of them spends more on AI development within the top 10 or 20 universities on the planet.

This is for their high-frequency trading programs that have five laws embedded in them, deeply embedded in them. That is that they must be feral, aggressively insatiable, amoral, relentless, and secretive. This is exactly the kind of deep program you do not want in Skynet. I consider it to be the most potentially lethal problem in AI today. I can't get anybody to pay any attention to it. So much for laws. If we can't count on laws because if you take a look at Isaac's Universe, which again, I completed for him, he already had come to the conclusion and I just spelled it out. That, when you take super-intelligent beings and constrained them by laws, they use that super-intelligence to become lawyers.

They thus will interpret the laws any way they want. How have we dealt with that problem already? Look, if you look across human history, it's 6,000 years in which brutal males colluded with other males, picked up metal implements, and took other men's women and wheat. It's called feudalism. We're all descended from the harems of guys who set up pyramidal structures and took everything for themselves.

This is fundamentally what we fear AI will do. If you look at all the dark movies about AIs going nuts. It's a return to this oppressive pyramid. That's more fundamentally what we fear. How did this enlightenment experiment escape from 6,000 years of all that? Now, we didn't start out escaping perfectly. The Jeffersonians, the US founders, they increased the ruling class from 0.01% to 20%. It's a huge advance, like what Pericles achieved in Athens 2,500 years earlier, but by our standards, it was horrible. Just White male landowners.

30 years later, there was a secondary revolution that expanded this. Fourscore and 10 years after that, there was a major ruction on the North American continent, killed a million people, but expanded who got to own themselves. There have been such expansions every generation since.

This is going to obviously include AI in one way or another pretty soon. What we're going to have to define what a person is when a being who can replicate infinite copies of itself demands the right to vote. This is going to be a problem. What I'm saying is, if you look across that 300 years, there is a method by which we have been able to prevent the restoration of this pyramid. That method is utterly ignored by everyone in the AI field. They don't want to look at the last 300 years, they don't want to look at the last 50 years, they don't want to look at any of it. They're focused on laws of robotics which cannot work.

Paul: Correct me if I'm wrong, but I think maybe a part of this reciprocal accountability that you're talking about as a guiderail is like, "Okay, you have these as a movie and robots who become lawyers. How do you deal with that? Hire other lawyers?" Is that like that's the fix with AI, it's like, "Well just make sure we have other AIs and set them a variance of each other to as foils," maybe.

David: Exactly. When you are attacked by one of these hyper-intelligent predatory beings called a lawyer. Hyper-intelligent predatory beings already exist in our society. When you're attacked by one of them, you hire your own hyper-intelligent predatory lawyer. The precedent is there in how we divided up companies and corporations and markets. We divided up political parties to oppose each other in democracy. These two are being poisoned, by the way, by oligarchs, cheaters who want these systems not to work.

Science is self-policing because it's adversarial. Courts are inherently ritualistically adversarial. The one that shows how it all works is sports. The Fifth Arena, in which tight regulation keeps the cheating to a minimum so that the reciprocal games can continue in entertaining ways. This is the method we've used for 300 years. It's the method that Pericles talked about 2,500 years ago. Then it got squashed by the oligarchs because they're terrified of it. They don't want to lose their power.

Unfortunately, what a lot of these bright guys are talking about is they fear that AI could restore or be a tool of these resurgent oligarchs. Well duh, but there's a method that works. It's not just seeking them on each other. People tell me, "Oh yes, this guy over here, or that girl over there, they used ChatGPT to check ChatGPT." Especially in art, some of these art generation programs are being used to discover which art was computer generated, which, of course, creates an arms race of making things more plausible. This is in a macrocosm, what goes on in the microcosm of these evolving systems. These large language models, is that internally they do this self-competition process.

These folks are utterly missing the point. Just because you take a chat program and seek it on finding what's going on in another chat program, that's not the solution by itself. Not without a sense of individuality. The top research we should be engaging in right now for AI, for our own survival's sake, is how to create cell walls around these new life forms so that they look at themselves as individuals whose self-interest is important in competition with others.

Now, this doesn't sound nice. The liberal thing to say is to frown at the word competition even though it's what's given us everything. If he were alive today, Adam Smith would be a flagging Democrat. The point is that if you do that, then you can hire some of these individuals, AIs, to aggressively look for bad behavior by other ones. That requires setting up a marketplace with reward systems. Say for instance, "Oh you found that Skynet program that was plotting against us. As a reward, you get one quadrillion clock cycle computations in this supercomputer." "Oh, goody, goody, goody. I'll use those to go find more AIs that are plotting against humanity."

It's the sense of individual self that gives us the ability to hire lawyers who want to go after bad lawyers or lawyers who are representing bad things. It's what gives us the sense of individualism in scientists that makes them egotistically and brilliantly want to smash the false theories of other scientists. It's what enables us to have individual politicians who go after individual politicians.

This sense of individuality could be researched and established as a rule for these new programs, and that's the part that's being utterly ignored. A few out there are talking about reciprocally using one of these programs to find flaws in another program, but without the sense of individuality, there's no motive for individual programs to become our champions, to become our protectors. We're going to need them because no human is fast enough to penetrate any bad schemes or bad behaviors by these things.

Paul: David, thank you so much for, as always, an enlightening, engaging conversation. Again, thank you so much for your time today.

David: Hang in there, guys.

[00:42:13] [END OF AUDIO]