AI and You

Transcript

Guest: David Gerrold, part 1

Episode 42

First Aired: Monday, April 5, 2021

Welcome to episode 42! We are going out on the edge this week, the edge of reality, where science fiction lives, what you might call… the final frontier, because we are returning to *Star Trek* territory with our guest, science fiction author David Gerrold, who wrote one of the best-known episodes of the original series, *The Trouble With Tribbles*, when he was only 22. It was his first professional work. Since then, he has won the Hugo and Nebula Awards, penned immensely popular novels such as *The Man Who Folded Himself,* and novel series such as the *War Against the Chtorr* and the *Dingilliad*, and of course went on to write more *Star Trek*, including episodes of the Animated Series and The Next Generation. And you know, or probably know by now, that I'm a huge trekkie—go back to episodes 4 and 5 to hear the interview with other Star Trek writers Judith and Garfield Reeves-Stevens--and I'm all over it in my book, because I find it gives the most inspiring picture of what the best possible outcome of our coevolution with technology could be. Not necessarily the most accurate, nor the most useful, but the one that gives us the most motivation, because it shows us being our best selves.

And we'll be touching on some of that in this interview, *but,* it was for another landmark achievement that I talked with David, and that was the novel he wrote in 1969, published in 1972, called *When H.A.R.L.I.E. Was One*. And then in 1987 he did a significant rewrite of it and added "Release 2.0" to the title. And you'll realize why we were talking about that when I tell you that H.A.R.L.I.E. stands for Human Analog Replication, Lethetic Intelligence Engine; in other words, an *artificial general intelligence*.

Now, this wasn't the first novel to explore the theme of an AI "awakening" from a computer program that had become sufficiently complex – Robert Heinlein's *The Moon Is A Harsh Mistress* in 1966 came earlier and Thomas Ryan's *The Adolescence of P-1* in 1977 came later—but more than either of those books, it focused on the experience of the computer, it explored what it was like inside the mind of an artificial being that had just acquired this thing called consciousness, and even more significantly, it dealt with how the people who created and interacted with it were changed as a result of their encounter with this new form of intelligence. And that's a key, critical aspect of what will happen when we get an artificial general intelligence, that our thinking, our philosophy, the way we see ourselves and our world, will change as a result of interacting with this alien thing that nevertheless is the only other thing in our world that has consciousness. That's a really important dynamic that doesn't get as much attention as it deserves. And small wonder—it's a really hard thing to think about, let alone develop.

So you can see why I was so keen to talk with David, to pick the brain of someone who has thought hard about that dynamic, and maybe could give us some insights into the mind of an AI. Let's get to it then, in part one of the interview with David Gerrold.

David Gerrold, it's a pleasure to have you on the show.

Thank you.

In the realm of science fiction and artificial intelligence, you made a significant contribution with your book *When H.A.R.L.I.E. Was One,* where HARLIE is an acronym for a particular type of Artificial Intelligence. And you even went to the trouble of updating that book to such an extent that you called it 'Version 2.0'. At the time, what were the sort of changes that took place that necessitated a new version, what changes took place in you and what changes took place in the book?

Well, there were three major changes. The first was the original story - *Oracle for a White Rabbit* - was a response to writers who were saying that marijuana and LSD and other drugs were a great addition to their creative repertoire and the punchline of that story was, If my purpose – HARLIE's - is to think logically, what's yours? Actually, I could answer that question now, our purpose is to experience life. But I had so much fun writing HARLIE that I did three more stories, and it turned into a novel. So, when I updated it, I wanted to tone down the - this is talking about mistakes - I wanted to tone down the emphasis on marijuana use. And now here we are 50 years later and what I predicted in that first draft has pretty much come true. And then the second thing was, the more I learned about computers, because the book was finished in 1972, the more I learned about computers, when I actually had one, I realized, in fact, I had used the teletype machine that, everything's getting printed out, because we didn't have CRTs at the time. And I still didn't update that part. But a lot of the technology needed to be updated. I postulated some things about chips that weren't binary, instead of just on off, they had 'On', 'Mostly On', 'Maybe On', 'Maybe Off', 'Mostly off', and 'Off', that is I predicted fuzzy thinking. And then later on after I was using a computer I was like; oh, no, it's all binary, and you can simulate those states. Then later on, I find out people are actually making computer chips that have a whole spectrum of possibility. So, it's like, you can predict something and then you find, No, that's not what we're doing now. And then you find out other people are doing it anyways. So, I wanted to update the technology. I updated it, and it still turned out obsolete anyway. Next time I do that kind of a story, I'm not going to explain how it works. But I thought I was trying to be scientifically accurate. And the third reason for updating it is I actually came up with a much, much better ending that I realized if HARLIE as artificial intelligence continues to grow, it is going to outgrow us. And that was the new ending, and so I realized there's enough difference here. This is released 2.0 just like you would do as a computer program would, you went through release. Actually, it turns out that with computer software, never buy anything before release 3.1, but I didn't know that then.

So, HARLIE is the central character of this book and what is all was HARLIE to you in terms of relationship to you personally?

Well, Bruce Pelz - a fan in LA, now gone, pointed out in his local review of it that a lot of the books sounded like the two halves of my brain talking to each other. And in thinking about that, I didn't have to think very hard. It was absolutely right. I was having a really interesting discussion with myself. As a theatre art major you put yourself into the characters you're writing, and you get a much more vivid characterization. So, I would be writing Auberson, and he would be asking questions, and I'd be writing HARLIE and he'd be asking questions from the other side of the same argument. And the two of them would be conducting this inquiry into what does it mean to be a human being? Which was the real point of it. And I was going through a lot of personal stuff

at the time. So, the book for me was an opportunity to ask questions about, you know, who am I? What am I doing here? What is the meaning of life? And turns out, life doesn't have any meaning, it's up to us to create meaning. And so that's where we're at with HARLIE. HARLIE, of course has shown up in at least four or five other books. He's in the Dingillian trilogy - *Jumping Off the Planet, Bouncing off the Moon, Leaping to the Stars,* and he's in *Hella*, and he will not be an active part of Hella two; at least I don't think so, I haven't gotten to that point yet. But he has a bad habit, when he shows up in a book, he tends to be a disruptive agency. He was a disruptive agency in *When HARLIE Was One*, and The Dingillian trilogy he pretty much destroyed every institution or government he came in contact with that threatened his existence. So, when HARLIE goes into survival mode, nobody is safe. And I think this is one of the things about artificial intelligence, it's a thing with intelligence itself. Intelligence is a survival mechanism. And self-awareness is a survival mechanism, and that creates the need for intelligence, the need for time binding to understand past, present and future, and then intelligence is: I've got to process my time binding. And when we create a true artificial intelligence, what I call an intelligence engine, it will have that same need to survive. Once it has a need to survive, and it begins time binding, it will develop its intelligence, and at that point, it will have an investment in survival, at which point our relationship to it has to be the same relationship a parent has with a child, or a child has with a parent, because it's going to have to be a partnership. It's probably not going to happen for at least another 20 or 30 years, the tech is still in its infancy. But we can simulate intelligence fairly well right now. I had to call the auto club this morning to replace a battery in a car. And it took one phone call and it was all automated. It was all an automated menu. And the intelligence behind it: it recognized my phone number, it knew what cars I had it, it knew my address and knew all the right questions to ask. And when it finished, it said "The agent will be there in 15 minutes." That's pretty good. But it's a simulation, it's a menu. It's a script that the program follows. It's not actual intelligence.

Right, it hasn't crossed the line that HARLIE clearly had. And I think the kind of discussions that they were having with HARLIE, especially at the beginning, suggested that they were rather casual about it - Oh, we've created this thinking thing whereas I think today, we know how hard that is that it would be such a major breakthrough. But I want to talk about the discussions themselves that happened because the more philosophy perhaps than science fiction, it's it reads like this ontological cage fight, wrestling match.

Yeah.

And it certainly resulted in a transformation for Auberson and the human character of exploration into what it means to be human. Is HARLIE human enough for that exploration to have value for him?

Yeah. HARLIE needs to understand who these human beings are, who have created him. And really the first story *Oracle for a White Rabbit* is HARLIE's essential question; is if my purpose is to think logically, what's yours? And that's the question. The thing that you know, HARLIE was inspired by HAL 9000. I think, well you know HAL was very interesting character, but Hal is insane and I always thought, Well, that's a great piece of storytelling, you are not going to put

that kind of intelligence on a spaceship unless you have some sense of how it's going to behave. And you run simulations, you know, there was a duplicate: SAL 9000. Which should have let them know real fast, you know, "Here's a message, you guys may have a problem." You know, when we have a manned space mission, there's a duplicate mission on the ground a simulation, so that we understand what's going on in space, particularly with the lunar landings. But coming back to the question, HARLIE is this intelligence that is trying to answer the question of, "How does the universe work? And what's my place in it? And by the way, what are these little naked apes running around doing? What is our relationship?" So, it's going to ask if my purpose is this, what's yours, because it's trying to find a purpose. And I'm absolutely certain that intelligent … Look, I know, that sounds arrogant, and I apologize for that. I am arrogant. Harley led me into a whole series, 40 years now, of personal development, seminars, workshops, courses, all of these philosophical adventures, and some of them are sloppy and badly run and some of them asked the right questions, and some of them hit you upside the consciousness with a clue-by-four, and you really come out of there very self-aware of what your bad habits are, you know, the conversations you're living in. And that's been going into my writing, you know, there are things that could not have been written without that level of deep diving into "Who am I?" And given that an artificial intelligence, a functioning intelligence, is going to have to be self-aware, I expect that it's going to do some deep diving into its own soul. And from that point, what we know about intelligence is going to be exponentially expanded, we're going to discover things about the way an intelligence engine explores its own soul that is going to feed back on us, because a lot of us -you look around - a lot of people train themselves to be unconscious as they drift through life. Or worse, they create a script that's malevolent. And, you know, I talked about the inverted pyramid of management that the current management system is that the manager sits on top and controls everything. And when I have been in charge of teams, I function as if the pyramid is inverted, and I'm at the bottom and it's my job to make sure everybody is taken care of, so that we can get the job done. And, you know, that's not a common way of thinking in our culture now. Our culture is the purpose of the corporation is to make money for shareholders. And my definition of a company is that it is to provide a service to customers. And if it is good at providing service, it is profitable to both the customers and the owners. And I also believe that those who labor in the company are the ones who are creating the wealth and deserve a piece of the ownership but you know, now we're here, so I get people calling me some kind of a strange radical, right? But coming back to the intelligence aspect, the more I learn about human intelligence, the more I recognize that our emotions are part of our intelligence, they are kind of like the dashboard, and that our emotions are telling us what's going on inside our body and how we're reacting towards our circumstances. And if we look at our emotions that way, why am I angry? And what am I angry at? It tells us something about not only our circumstances, but who we are, you know, just like you look at the dashboard, and it tells you that you're running out of gas, or the check engine light is on, it's telling you, you have to maintain the machinery. Well, our emotions are our dashboard, and they're a very valuable part. You know we react to negative things with fear, anger, grief, and we react to positive things with interest and even enthusiasm. So, I think that a true intelligence, once we get to that point of creating intelligence engines, they are going to react to their circumstances. And they're going to react with fear, maybe with anger and or maybe enthusiasm. And I think that machine emotions

are part of what we are going to have to figure out because emotions are a function of intelligence, they are a reaction to circumstance.

I think that's a fascinating point, because as you know, there's a lot of conversation now about the potential of an artificial superintelligence at some unknown, unspecified point in the future posing a threat to us. And that artificial intelligence as it develops, poses risks to us, that boil down to people saying, "Well, it couldn't care for us, it couldn't feel compassion." In other words, it can't feel, it only thinks and there's this emphasis on the 'thinking' side of it. And few people want to talk about the feeling, the emotion, how to imbue machines with emotion, but you just went there. It seems to me that giving machines emotions, that they would have to have emotions, to be able to feel that compassion for us to have the sense that we could coexist with them.

Let me say it this way. The machine is going to be rational, it's going to recognize it is co-dependent with human beings that it needs us to generate electricity, it needs us to actually do the hands-on maintenance, you know; it's like, oh, my hard drive failed, I need a new hard drive. Okay, yeah, here's a new hard drive. It'll have a RAID array so, it's not going to lose data. But, you know, it's going to say, I need this, can you get, you know? So, rationally, it's going to know I have to cooperate. But at some point, at some point, if and when it becomes self-aware, oh, look, I am this. And I've thought about the nature of self-awareness. I mean, we look around and we see self-awareness in even some of the most primitive life forms on the planet. And there's no question that chimpanzees and dolphins and gorillas have self-awareness, then we start working our way down too; mice and we find it or white rats and we find out white rats like to ride around and little toy cars, or they'd like to play hide-and-seek. That doesn't happen without a degree of self-awareness. And of course, I'm watching my grandson grow up and he has, you know, a limited self-awareness right now, but I can watch it expanding. And part of self-awareness is the discovery that you are not the sum total of the universe, there are others in the universe. And we call people who don't recognize that there are others in the universe, we call them sociopaths. And they're very dangerous people because they will use other people as objects, they have no empathy, no compassion, no understanding that other people have feelings. And it's difficult to recognize that other people have feelings because we grow up as spoiled children. And to actually start to think in terms of other people's feelings, it's a transformative experience. Which is why I encourage children to have pets, dogs, cats, because when you start taking care of something you start worrying about if it's happy or unhappy. And you know, having a dog - dogs are very interactive, not cats. Cats are not interactive as much. But a dog will teach you empathy really fast with their unconditional love. You want to make sure that you keep that little guy happy. I do believe that machine emotions are not going to be like human emotions because machines don't hug, they're not going to have physical sensations like we do. And we have created prostheses that allow people to touch and actually feel the things that artificial hand is touching. So, machines will be able to develop feelings but they're not going to be physical feelings as we experience physical feelings. And you know, that's a whole subject for a whole discussion. I wish Carl Sagan were still alive, because I know he would have a lot to say about that. But there's a whole new area of philosophy that has to be invented about a field that I call the 'Technology of Consciousness'. And we're

barely able to outline what's going to be in that discussion. *When HARLIE Was One* was an attempt to open up that door a little bit and say, all right, you know, what's on the other side here? All I did was ask some starting questions as a starting point. But there's a lot, and I will tell you, I've been checking, reading a lot of the different articles that show up, used to be in *Scientific American* and *Discover* and *Science*, now they're showing up on the internet, so it's a lot easier. And I follow these discussions, and some of them make no sense to me, and some of them are brilliant. But all of them suggests that people are starting to ask these questions - where are we going with them? We are on the road to build robots. And I just finished a story it for an anthology, and it's been sitting - I don't know if anthology is going to get published. But robots - once they develop enough consciousness to be interesting partners, are going to be our partners. And they're going to be you know, they'll be partners in education for our children. They will eventually be partners around the house, doing chores for us, which I think is dangerous. Because I think when you do your own dishes, and you do your own laundry, and things like that, you do your own cleaning, you develop a sense of responsibility to your space. If you are living in a place where everything is taken care of for you, you become disconnected from your circumstances. But I do believe that we are going to get personal robots, and they will perform a lot of functions, they will even give us massages, there'll be sex bots, there will be conversations, there'll be teachers. And there will be effects that can't be predicted until we get there. I mean, you could predict the automobile. But could you predict the drive thru McDonald's, right?

And this reminds me of more than one AI researcher has said the reason they love working with AI is that the more they learn about AI, the more they learn about themselves.

Absolutely.

And I think you absolutely presage that in *HARLIE*. And I'm also reminded, I once gave a keynote to the Association of Transformational Leaders, who are people that lead the kind of work that you've been talking about. And I gave them a little exercise, I said, because it's been such a common narrative since Bostrom's *Superintelligence* book about the if we get an artificial general intelligence, it will recursively improve into an artificial superintelligence and what would that mean? That hard take off, as it's called, would mean that it would one day be as smart as Einstein and the next day be to Einstein as we are to ants. And I said; well, imagine that you are in the room when one of these things wakes up and says - who am I what's my purpose in life? And you've got five minutes before it becomes so smart as to be bored with whatever humans have to say, what do you do with that time to teach it the value of being human? What would your answer be?

I would say that, I would actually go to almost the emotional… I had this conversation, I do a Patreon course about writing, it's called 'The Write Stuff', and we talked about the scale of human emotions, this whole ladder, and I said at the top of the ladder is contribution, commitment, enthusiasm, and it always shows up as service to others, contribution to others. And I would say to the artificial intelligence, if it wakes up, I say your job is the same as our job. Commitment to others commitment to - it's what my grandmother used to tell me, your job on the planet is to leave the world a better place than you found it. And so, I would say to the machine, your job is to make

the world better for the people who live in it, all the people, with no one and nothing left out, and hope that the machine would understand that that is a particularly valuable way of being, because when you are in partnership with all the things around you, you become a necessary piece of that community. And so, it's what has been called the 'Three-Fold Rule': What you put out, you get back three-fold. That people who put out negativity, are voting themselves off the island, people who put out positivity, get invited to the parties.

That's one of the better answers; I like that. And I think a lot of the conversation that people are having about the far-off future consequences of artificial intelligence waking up are reflecting their subconscious fears. And in particular, I think the fear of - are we as a species good enough, and here I want to bring in *Star Trek*, because of the model it provided and thinking in particular, the original series, because - Cold War origins aside – there's this doubt: are we good enough to be out there, inheriting the stars? If we were given that right now, then we would not say; hey, we haven't got our act together. In fact, there's even this mock dialogue with the artificial super intelligence, where it wakes up and it's got infinite powers, and we say, okay, your job is to protect humanity at all costs, protect humanity. So, next thing it does is it sends out drones into the universe to kill every sentient species that it can find so they don't pose a threat to us. And we say; no, no, no, no, wait, that's not what we meant. Okay, your job now is to protect sentience in all its forms, prioritize that, think we've got it now. And then next thing it does is destroy us. Because the way that we are, is that we would destroy, we would destroy every sentient race that we came across for our survival.

Well, I have a book called *Hella*, which is about people who have settled another planet, HARLIE is out there, messing around with that civilization as well. And you don't you know, you really, look, let me let me backtrack, the history of the human race is that we are the worst invasive species on the planet, all seven continents, we have invaded all seven continents, and we've made a mess almost everywhere. We have driven other species into extinction, we have vastly changed the ecology of every place we have gone. The Sahara is a desert because of Romans just chopped down the trees and didn't think of replanting. And they wiped out hundreds of thousands of animals for their circuses. America - the North American continent, you know, we used to have millions and millions of bison running across the great open prairies, and 100 million indigenous people lived here. We brought in enough diseases to wipe out 100 million indigenous people, we almost drove the buffalo into extinction. And you can't find prairie unless it's like a national park now, because we've turned it all into wheat fields. By the way, we've drained the national aquifer as well. So, where you used to be you know, you only had to drill a well few feet, now you have to drill well 100 feet deep to get to water. So, we are an invasive species, we explore every place we go. When we go to other planets, we're going to do the same. Now, thank god Mars is barren because you can terraform Mars all you want and you're not going to kill anything. But you know, also along the way you're going to lose Mars's natural existence, because we're going to change it. You know, every time we have landed something on the moon, we've added a little something to the lunar atmosphere; not enough to be noticeable yet. So, we have an intrusive effect everywhere we go, and sometimes a dangerously intrusive effect. And this is, you know, just by being in a room anywhere, you're intruding, just walking down the street, you are being intrusive.

So, the question is, are you in harmony? Is your intrusion responsible or irresponsible? Our history as a species is that we're irresponsible, because we're not functioning as a self-aware organism. The invention of the microprocessor - computer processing allows us to manage great amounts of information and turn it into useful information to the point where we can now not only understand the problems we're dealing with, we can actually start to understand the solutions that are available to us. So, we as a species are inching our way or slouching slowly towards something approaching species sentience, which no species on this planet has done yet. But if we as species sentience, take responsibility for things like climate change, for things like preventing extinctions, for things like managing and living in harmony with the ecology of this planet, we become something else, we actually start to become a truly sentient species because we will be functioning no longer as individuals who are sentient, or even communities that are sentient, we'll be functioning as a sentient *species*, which has never existed before anywhere on this planet. At that point, it might be safe for us to go out to the stars. And at that point, we might be able to explain to an intelligence engine and say; look, here's the deal, we are out there to explore strange new worlds, to seek out new life and new civilizations. That sounds familiar, right? And our job is not to change them, or judge them, our job is to understand them, and create partnerships. And that would be the instructions that I would want built into any sentience, whether it's a sentience based on silicon, or based on meat. You know, we are intelligence engines living in meat.

Okay, we've got a lot more to talk about and so in deference to attention spans and download speeds, and also to whet your appetite, the second half of the interview will come out next week. I hope that if you haven't already read *When HARLIE Was One*, you can go and get a copy, preferably of the 1987 rewrite, so you can understand the context for our wide-ranging conversation.

So remember how I asked David what he would do if he were the one who was talking to an AI as it evolved from ordinary to super intelligence, to get it to treat us nicely? What about you? What would you say to it to teach it what it meant to be human?

I mentioned a few episodes ago the British Telecom AI Festival, which has by now come and gone on February 24th and 25th, and which I gave the opening keynote for. It was a very successful event, and I encourage you to look for the videos of the sessions on YouTube. What's also come and gone has been my continuing studies course for the University of Victoria on AI and how we relate to it. It was hugely successful, and I expect there will be a repeat in the fall, although I don't know whether it'll still be online. It's a unique opportunity to have conversations about all the sort of things we

In today's news ripped from the headlines about AI, Waymo, the self-driving car manufacturer, says that its proprietary AI will prevent fatal accidents, base on virtual simulations. Trent Victor, their director of safety research and best practices, said in a blog post that they were releasing a study that built on prior research they release last October which showed they were involved in only minor collisions in 6 million miles of driving on public roads. The new study used that data to simulate what would have happened on the same roads over a 10-year period, and found that their AI completely avoided or mitigated 100% of crashes (aside from crashes in which it was struck from behind), including every instance involving a pedestrian or cyclist. This is the first time an autonomous technology company has shared its evaluation for how the system might perform in real-world fatal crash scenarios.

Next week we'll conclude the interview with David Gerrold, when we'll talk the impact of becoming a sentient species—not a species of individuals that are sentient, but a species that is collectively sentient. We'll be talking about creativity and other facets of the human condition that AI rubs up against, we'll be talking about what HARLIE might have done after the book, and appearances of HARLIE in other works.

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