# AI and You

Transcript

## Guest: Kenneth Stanley and Joel Lehman, part 2

## Episode 136

First Aired: Monday, January 23, 2023

Hello, and welcome to episode 136! Today we will conclude the interview with Kenneth Stanley and Joel Lehman, authors of a unique book in the field of AI, titled *Why Greatness Cannot Be Planned: The Myth of the Objective*. As we said last week, although this book is shelved under Computer Science, it sounds like it's clearly a management and leadership type of book, and in fact it is. What's fascinating is how what started as apparently a straightforward experiment into using a genetic algorithm turned into a lesson in psychology that's applicable across so much of human experience and when Ken and Joel realized what that was, they got quite excited about bringing that message to a wide audience.

Ken Stanley was previously Charles Millican Professor of Computer Science at the University of Central Florida and was also a co-founder of Geometric Intelligence Inc.

Joel Lehman is a machine learning researcher interested in algorithmic creativity, AI safety, artificial life, and intersections of AI with philosophy and psychology.

Both Ken and Joel were at Uber AI Labs, where Ken was head of Core AI research and Joel was a founding member, and they were both again at OpenAI, co-leading the Open-Endedness team (studying algorithms that can innovate endlessly).

In this conclusion of the interview, we'll be getting much more into the philosophy that AI surfaced for them and how to apply that in an organization or your life. Here we go.

**Peter:** So, we've seen some alternative organizational structures emerge more, in recent years, I'm thinking of things like holacracy, where there is no hierarchy, no apparent leader, and they might be able to take on board a philosophy like this more. So, if there was a place which had the resources and the inclination to say, "We want to make this work," how should they do that? What are the practical ways of realizing your discovery?

**Ken:** Well one thing is to separate the kinds of objectives as Joe described them, that are within the normal range, or what I would say like not super ambitious, from those that are ambitious and oriented towards discovery and innovation. Because the normal types, you can do those. So if you take this as a sort of a message, which is over the radicalized, it's like, forget all your objectives and just do whatever you whatever you find interesting at all times, like, of course, this will cause organizations to fall apart. We need to keep doing some of the things that are normal. But when there are components of an organization that are oriented towards discovery and innovation, which I think is essential, especially for large organizations, because like that's how you grow and prevent corruption and things like that is to innovate, those organizations can stand to change, and probably should, what you should do is you should focus more on the accumulation and collection of stepping stones, and also on honoring what your employees think

I think is interesting, which is a very nonobjective thing. Because if somebody says, I want to work on this, because it's interesting, there's a very few places in the world where you're going to get validation for that. It's like, what's your goal? Like, what are we going to get, like, when how's it going to fit that hit the bottom line, so we go back to objective thinking. But what I'm saying is that for those things, where we're trying to innovate, we need to allow people to do things that they find interesting and we need to accumulate stepping stones, which means that like in the PICbreeder spirit, the more things you find that are interesting, the more things you can find that are interesting and so the ability to sort of lock those in remember those that people can then later be exposed to them and build on them, if they find them interesting, is important in order to have innovation and those things, especially this this idea of embracing interestingness, as a valid topic, are very difficult, because it's thought of as subjective and it gets us to the counterobjective and so you think of that as sort of dangerous, because we're afraid of things that are subjective. But what you have to remember, like the reason it's justified is that it's not just we're asking some random person on the street what's interesting, like, that doesn't make any sense. So, just go out and say, "hey, you think we should do this?" Oh, yeah, that sounds interesting. No, we're talking about experts. You hired these people because they're supposedly experts in this domain. How can we deny that they have any instincts about what's interesting? If they completely lack instinct about that, then I would argue they shouldn't be considered experts in the domain. But if you do think you have experts, then their instinct for the interesting is your most valuable asset if your purse doing innovation and to deny it and say, oh, no no no, we can't talk about what's interesting because we have to have some performance measure. Because we have to be objective, you're denying the best talents that those people have. And so that's where I think there's room for some significant change.

**Joel:** Maybe to add on that just a little practical element on the ground would be to make sure that there's not an overpowering drive for consensus early on, in that process of discovery so that if there's one leader who is kind of organizing people, and is kind of averaging the ideas of those people beneath them or something, then that will stand as an obstacle and I think that it's kind of somewhat common for that to happen. So not only should there be an appreciation of what's interesting, but also an attempt to cultivate an environment of disconsensus in the end organs of an organization that you want to be more geared towards innovation, or your point.

**Peter:** It's like you've established a formal basis for the justification of recognizing intuition. Because when someone says, "I want to do this because it's interesting", you're saying, that's enough reason for doing that and that if they actually had a formal justification that was recognized as being a way of getting towards the goal of the organization, it would not fall within the structure that you're promoting. You're saying that it's in the cases where you *can't* come up with such a formal justification, just saying, "it's interesting, but I can't tell you why," that this, the gold of what you've discovered lies.

**Ken:** There's a lot of security blankets floating through this because it's like all these things about formal justification is really about security blankets, you just want to feel like the things are not going to go off the rails and that but one thing to think about. I mean, by the way, just because you have a security blanket doesn't mean you're secure. So this is a big problem. But the but by the way, this thing, what's important, also to keep in mind is that we're not saying that, like you just you're justified *just* because you think something's interesting. Like that would be a fairly dangerous thing to say, like anyone who thinks something's interesting, should be getting funding to do that. Of course not, that would be silly. But what the real point, I think, is that we should be able to discuss what's interesting. If I'm an expert on something, and you're an expert on something, and I want to do something, say, you're my boss, and I say, this is interesting and I want to do this. Well, of course, I should be able to defend that, like, why is it interesting, I should be able to talk about it and I don't, we're not at all saying you can't have a discussion about what's interesting. But the point is that *that* discussion is a valid discussion, even though it's independent of the performance measures of the objectives and the security blankets, it's a valid discussion, because both people are experts in the subject matter can actually engage in this discussion. It's also important to about like consensus, you don't want to make it to everybody has to agree that because that will cut off all these stepping stones, people won't pursue these directions. So, we need to be careful to let a number of diverse positions move in different directions.

**Peter:** And now you've been talking about leadership and team building and organizational culture and that's an incredibly active competitive field in consulting and sections, entire cases, in the bookstores. Have you explored that as consultants, have you set up shop to bring that message, because the passion that you're evidencing here for that lies to me squarely in that domain that I could see you in front of any number of boardrooms, saying to a receptive audience, here's how to embrace this for 10x kind of change.

**Joel:** I guess that's not something that I've at least explored myself and I guess I would be excited for someone to do that. But I think that's not where my joy in particular lies, but it does seem like an important impactful thing that could happen.

**Ken:** I've been going on I'm talking about this for years, but not as a business consultant. I mean, I've talked to groups when they wanted to talk about it, and individuals and I get emails, like people just say, I've been thinking about this, I want to think about what this means. But yeah, I haven't done it on like a formal paid basis or even tried to do that and I guess, yeah, I think probably like, Joel isn't really, my heart isn't really in being a consultant or business business guru type of person and yet, I do feel passion, obviously, for the fundamental insight here, like, I feel like it's very important to get it out there. So, I feel more like I'm just trying to spread awareness of it and to provoke a discussion than to be a formal consultant, who would like to go to boards and sort of get paid positions to do this. But I agree with Joel that it's interesting. Because it probably is worth doing, if somebody would take up that mantle. It would it would be a role that I think would be would be provocative and probably productive for the world. So, it's just a question. Like, personal- our personal, various push and pull of our priorities are in our live. It's so foreign, I come out of like scientific research to just go and become a business guru. So, it's hard to understand that a straddle that or what I should really do.

**Peter:** Well, I'm just voicing the impression I get which is, especially being in San Francisco, you could walk around the corner and find a boardroom of a startup that would listen to you talk the way you just have for the last 10 minutes, and be engaged and motivated with that. Recognize that maybe that's not the direction you want to go. But it's certainly a possibility. And so we've been talking about something here where I like to say that AI holds up a mirror to us, and you've demonstrated how that has pulled you into the realm of philosophy and psychology and really opened up a new domain to you. Have you seen that do that with others? Do you have students that have come on board and got excited and positions started to take this in new directions?

**Ken:** we see within the field, for sure, like people have taken the mantle of novelty search and built on that, like now there's a whole area, which is being called quality diversity, which is like algorithms that are built on these ideas. So you see that within research, I think you're making a point, broader point, which I think is really interesting is the meta issue about, like the mirror, I find that really fascinating, like the idea of artificial intelligence holding a mirror to ourselves. Because it gets, it gets a lot less of the excited attention that AI is getting right now, even though I find it like really important as a side effect of artificial intelligence is, what does it reveal about our own selves. And you would think that in studying or trying to sort of recapitulate the processes that go on in our own heads in something that's artificial, that we would actually reveal something about what goes on in our own heads and hopefully, if we're actually learning something that's actually important it would be at a deep level, you know, you would think that. And I think it's really interesting how little of that we see, it's just like, a lot of the kind of advances in artificial intelligence, they're revolutionary, it's like we just beat the world Go champion, that's a big thing. But you don't get the kind of like, corollary about like, wow, we really understand ourselves better now, this is who we are. You get some debates about like, well, well, that was actually hard. After all, like that always is like a debate like, actually, it's not really hard. That's not the hard part of being human. But you don't get much more than that. And so like, this is an anomalous event, I think with the novelty search, and the kind of discoveries for Picbreeder were really drove deep into like, what does it mean to be human? Like, what is it that actually leads to the most interesting things that we do as people through our intelligence, and to me, that's like what we should expect. That is one of the greatest things that AI can deliver and it's intriguing how little we talk about it from that angle, and I'm not exactly sure why that is. But it's a great meta point and I think that it is inspiring to people like students, like when students see that, they I do think they get excited. This is like not just a science about computers, this is like about humanity and of course, humanity is more important than computers. That's what the thing [is] that ultimately matters. So, I think it's very inspiring and worth drilling more like why we see less of that in the field.

**Peter:** Wow that was very inspiring, and passionate. And you talk about the application of this principle to AI because there is a field, which has a goal that we don't know how to reach, which is artificial general intelligence, and a lot of the approach right now is incremental, exactly the sort of thing that you would say is counterproductive, of "Let's throw another billion parameters at this and see what happens." Have you seen opportunities taken or opportunities missed since your book was written in the field of AI, in terms of advancing towards novel goals and capabilities of artificial intelligence?

**Joel:** Well, I mean, I can kind of skirt around that point. In that it's interesting how that the history of the field of AI is also littered with examples of deception and that kind of principle in that the neural networks themselves have gone through several phases of boom and bust and so there is a way to Ken's point that it's interesting how people who study the nature of search can kind of be misled by the same way that they're searching the space themselves. So, just that right now, it's hard to do anything but deep learning research. A lot of the funding is going that direction, that's where all the excitement is. And deep learning research is having a huge moment right now. And so you could say that one thing that's unclear about deep learning is where we are in the stepping stones relative to AGI and some people are quite confident that we're maybe just a couple of stepping stones away that we can just push on this paradigm and get there. And that may be true. And yet it seems like also, we should be mindful of the lessons of the past, and that we could be a couple stepping stones away and that we should still continue as a field to search broadly. So, it does seem like it's kind of a stifling field to be in some ways right now, just in terms of how many people are in the field and how much research is getting done and how much chasing there is of benchmarks and that kind of thing and it is like just striking like that. In theory, people who study AI are masters of understanding search and yet, there's a disconnect in the actual practice on like, one level up, and somehow translating those insights from the algorithmic level, just one level up seems somehow beguiling.

**Peter:** We've had researchers on this program excoriate AI companies for playing it safe and just trying to throw more money and electricity at deep learning instead of taking it in novel directions. On the other hand, each of you both of you were until recently at OpenAI, which has since come out with the ChatGPT model that is taking the world by storm and doing things that are novel by any definition of the word. And, have you any reaction to where that's going in terms of its novelty, in terms of its surprise, in the field of AI? It doesn't seem incremental.

**Ken:** I mean, it's obviously interesting and that's sort of like the word that we come upon over and over again, if something's interesting it probably should be pursued. This is interesting, it deserves investment. And, "does it lead to AGI" is to me a totally different question, than is it interesting that we could debate about will this actually lead to AGI in the sense of like, just a direct path? Like just keep scaling keep pouring money into it? Eventually, you have AGI like, that's a question. But it's still even if it's not the case, it's still interesting, and it's going to lead somewhere interesting. So, I think it's clearly at least interesting and should continue to be explored and then we can talk about what's going to lead to AGI also, which is a whole other thing.

**Peter:** In what ways have people most commonly misunderstood your message and what do you need to say to correct that misunderstanding?

**Ken:** Well, I think that one very common misunderstanding is that we're advocating acting randomly. Like a lot of the time, like the first thing that people say after they hear that, well having objectives can be deleterious to achievement, it's like, "well, what are you telling me, I should just do things that are random?" And I guess I'd like to correct that misunderstanding, this is not advocating randomness of being random is not a good policy, either. It's probably bad for you and so it's advocating using your entire life experience, which is anything but random, to think about what's interesting right now. And interestingness is not random either, because it is ultimately a comparison between what you're considering right now, and where you've been in the past, and that comparison is full of information. It's information-rich. In fact, I would argue it's richer than the information that you have about the comparison between where you are now and where you want to go, which is what we call your objective, because that's out there in the future in a fog, where we don't even know what it's like. But we *do* know about the past and we can look at that as a context and think about what that means for whether unknowns are actually interesting with respect to what I know about the world from everything that I've ever seen. That's a completely principled and totally nonrandom thing to do. The fact that people even think that is something like it's similar to doing things randomly, just shows you how completely oversaturated we are in objective thinking. It's just been totally shoved down our throats to the point that we don't question it at all. We think the only other thing you could do in life is just be completely random. It's absolutely not true.

**Peter:** You know, it feels to me as though this is actually now got an element of a cultural reaction against a common Western, predominantly American, and predominantly Bay Area startup - if I dare go that specific - culture of ruthless obsession with goals, and that your message might resonate more in other parts of the world that are less hyperobsessive about goals, does that land?

**Ken:** I mean, people say like an Eastern culture that is more of a kind of a Zen or something, and not necessarily objectively driven. I mean, I'm sure I'm caricaturing it, because I'm not an expert on other cultures. So, I don't want to kind of give a ridiculous caricature or something. But I think there's, there seems to be an element of that, you know, we even put quotes in from a Chinese philosopher in the book and that has been within some other cultures, you know, kind of an element of you know, the road is what's important and not the destination. Like you hear that. I mean, even in Western culture, some people have said things like this too, though. So, I think there's some truth to that, and is interesting how other cultures would absorb the message of the book like very different places than but where it's generally so far been circulating.

**Joel:** It's interesting, I think, also, that even within their culture, there're pockets of real deep appreciation for the message. I think, like, we think about venture capital, for example, where you might invest in a person maybe specifically because it's interesting, they're interesting, the idea is interesting, you don't know where it's going to go. So, I think like, there are avenues of even within kind of a culture that's maybe KPI-oriented, or lots of goals and objectives that there's a fit someplace within there.

Wow, well, one of our goals is to stay on time and - so that's an easy one to understand - and here we are, at the end of it, I will personally never read in computer science, the words "objective function" the same way again after this. How can our listeners, beyond your book, find out more about what you're doing and your message and where should they go to get that?

**Ken:** Well, I have a website. I think Joel has a website. There's one place you can go. I mean, we've published a lot, both of us and so you can find our publications if you really want to get into the scientific underpinnings of it. You can also find at least, there's a lot of videos discussing these issues like so you look up the name of the book, for example, on YouTube, *Why Greatness Cannot be Planned*, you can find shorter distillations of these issues in discussions and things like that. And I'm going to try to start a company. That's what I'm trying to do right now. I should let Joel say that he's going to do what he is doing.

**Joel:** Still exploring.

**Peter:** Thank you, Ken Stanley, Joel Lehman. Thank you very much for coming on AI and You.

**Joel:** Thank you for having us.

**Ken:** They were great questions.

That's the end of the interview. Ken and Joel are great examples of the sort of guest I love to have on the show because they're not here to sell or promote something, you can hear the passion that's driving them, and in their case, it wasn't so much to develop something with AI or to discover something about AI, it was a life lesson that AI revealed to them so that they could communicate about it to others. As I said, the metaphor that I use is AI holding up a mirror to us, and as overblown as that may sound, it resonates deeply with many people, including Joel and Ken.

In today's news ripped from the headlines about AI, NIST, the National Institute of Standards and Technology, has developed a scoring system that quantifies human trust in AI systems. The system involves two scores: first, a user trust potential score, which measures details about a person using an AI system, including their age, gender, cultural beliefs, and experience with other AI systems. And the perceived system trustworthiness score, which covers technical aspects, such as whether an outdated user interface makes people question the trustworthiness of an AI system. The proposed system score assigns weights to nine characteristics, including accuracy and explainability. Other factors and weights for factors that play into trusting AI, such as reliability and security, are still being worked out. The paper, by Brian Stanton and Theodore Jensen, concludes, "Like any other human cognitive process, trust is complex and highly contextual, but by researching these trust factors we stand to enable use and acceptance of this promising technology by large parts of the population."

Okay, now wake up, because next week we will be talking about consciousness, with none other than Anil Seth, author of the bestseller and 2021 Book of the Year "Being You: A New Science of Consciousness." Anil is a TED speaker with 13 million views and professor of Cognitive and Computational Neuroscience at the University of Sussex. 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**