SCIENCE OF THE NOOSPHERE

Stuart Russell and Terrence Deacon

with

David Sloan Wilson

David Sloan Wilson: Okay. So welcome gentlemen, I'm so excited to have this conversation on artificial intelligence and the Noosphere, Stuart Russell, and Terry Deacon. And let's have you introduce yourself very briefly to our audience, as to your disciplinary background. Stuart, starting with you, please.

Stuart Russell: Sure. So I'm a professor of computer science at Berkeley. I've been here, this is my 36th year of teaching. And I've worked in more or less all areas of artificial intelligence, with a primary focus, I would say, on learning and decision making, probabilistic reasoning or reasoning under uncertainty. And also some of the fundamental philosophical foundations of the field, what is rationality, how should we design AI systems to be beneficial to human beings and questions like that.

DSW: Awesome. Great. How about you Terry?

Terrence Deacon: So I'm also a professor here at University of California, Berkeley. I am biological anthropology and my other hat is in cognitive and brain sciences. My background has been mostly looking at the evolution of nervous systems, evolution of vertebrate nervous systems in particular, with a focus on humans and to some extent, on language. And I've become very much interested in issues of cognition and how the nervous system works and how the nervous system is similar to and different from the devices we create to do intelligent processes. So Stuart and I have that overlap and come from opposite sides of the question. I've been here at Berkeley since 2002, before that I was at Harvard and Harvard Medical School, where most of my work was in fetal neural transplantation, actually. So quite a different field, but as you can see, there's some interesting overlap here.

DSW: So what a perfect combination to talk about Pierre Teilhard de Chardin's concept of the Noosphere. He himself began with the origin of our species and then ran it all the way up to the future, including technology basically, he has a huge interest in technology, he was quite prophetic in terms of where technology was going, way back in the first half of the 20th century. And the expansion of the Noosphere, which is like this thinking envelope of human society, larger and larger scales, ultimately encompassing the entire earth, which of course could not take place without technology. And so regardless of how much you know about Teilhard, both of you are experts in the topic that he was addressing. And so Terry, why don't you begin by just outlining the map here, especially with respect to technology and the future, with the goal of global cooperation in mind, global governance and cooperation in mind, lay it out for us, and then Stuart can add to that.

TD: All right. Well, so one of the big issues of course, about being human is that we are dependent on each other for cognition. That is, even though we have nervous systems like chimpanzees, and in fact, there's no real fundamentally new parts to human brains, nevertheless, all of our cognitive capacities depend upon being embedded in this culture, held together by language and other forms of communication. So that in a large respect, we are not islands of cognition, we're this linked cognition.

DSW: I just wanted to insert language is itself a technology, and so we don't have to be talking about hardware. Language, spoken language is one of the first technological inventions of our species. So anyhow, now I promise I will not interrupt you, so please continue.

TD: All right. So the issue is that, of course, it's a technology, both of thinking and of communication, and communication in some respects is a kind of thinking, that it's a collective mental process. But in any case, what Teilhard realized back in the 20s and 30s when he was writing, was that in effect the growth

of technologies and particularly communicative technologies and knowledge technologies in the world, is bringing us human beings closer and closer together, making us more and more interdependent, making human thought more and more a collective phenomenon. And so this analogy of the Noosphere as a thinking sphere analogous to the way the biosphere is a sphere of life on the earth, for Teilhard what it meant was that in effect over the course of time, and particularly up till the present and now into the future, there would be more and more convergence of human thought. And so his notion of the Noosphere was that over time, this convergence would grow and grow and grow. And we're very much interested in that trajectory, in this discussion.

The key innovation that's happened really just within the past few decades, of course, as a result of computer technology and communications, added to that, artificial intelligence that is offloading not just work, physical work, but now cognitive work onto devices, is intermingled with and entangled with this process of human communication playing more and more of a role. And since Stuart is an expert in thinking about the future of AI, what AI has done, what it could do and where it's going, it's appropriate to pursue this question about what it's going to be doing, not just in and of itself as the technology, but what it's doing to us as now being in a sense cyborgs within that technology. So that our thinking process is not just distributed by language between us, but is now mediated by this elaborate web of communication that is amplifying this process by orders of magnitude.

My particular interest in conversing with Stuart about this, is not just to understand how that works, but also to understand the limits and the potential problems, because Stuart has played a great deal of the role of worrying about the control issue, how do we control this? And in particular, how do we control it when we are using it for means that might actually be destructive, for example, in terms of weaponry, in terms of sales, in terms of controlling people's thinking processes and so on. That's my hope that we can get to that Stuart.

SR: Okay. Well, so there's a lot of different topics, different roads we could take in this conversation, so I think I'll just mention a few of them. So one of them is a topic that I think pervades much of my recent thinking about AI and its role, and that's what human preferences are. And so version zero of the theory is that roughly speaking, humans do have preferences about the future. There are futures we want to avoid, such as extinction and enslavement and various other dystopias, and futures that we would like to bring about.

And this concept of the Noosphere is really important to that because we're not born with these preferences, they result from our immersion in the Noosphere. And so understanding the dynamics of that is extremely important because to some extent, our preferences about the future end up determining what future we get. We may not get the one we actually prefer, but our preferences control the way we act and one way or another, that ends up controlling the future. So that's the first point. I think another interesting point about the Noosphere viewed as this global thought process, as you put it, Terry, is that at the moment, I don't think AI is directly contributing as additional thinkers in the Noosphere, but it's dramatically affecting the connectivity of the thought processes that happen in the Noosphere, and it's doing that primarily through recommended systems operating in social media. And it's interesting to think that Stalin and Kim Il Sung and Pol Pot had far less control over their subject's cognitive intake than Facebook and Twitter do today.

They control the cognitive intake of billions of people, for hours and hours a day, and probably the majority of their non familial and non-work related cognitive intake. And so that's incredibly important and we have absolutely no idea what effect that's going to have, but it's going to have an enormous effect. One way of thinking about it is like, think about material science and what makes something a magnet. It's to do with lining up all of the spins of the atoms in the magnetic material, and to some extent what AI could be doing, what these recommended systems could be doing, is creating these magnetic domains where all the spins are lined up in the same direction because of the mass effect of

running the same algorithm, billions of copies of the same algorithm affecting people all over the world, and that can't do anything but have a huge effect on how the Noosphere operates.

Just as when you take a normal material and turn it into a magnet, it's now a completely different thing and has all these new macroscopic properties that it didn't have before. So that's a real concern, I think, and we're just getting to grips with acknowledging that it's already happened and the revelations from Facebook and others saying, "Oh yeah, we knew that we were ripping apart societies and destroying democracy and creating these huge islands of completely misinformed people. But we were making so much money and having so much fun, that we just carried on doing it." So this, we really have to get to grips with.

But then most of my recent work on thinking about the implications of AI is not about current AI, which I think is actually still quite rudimentary, and I think we're probably going to go through several more boom and bust cycles in AI, because I think there's way too much hype going on right now, just as there was in the 80s with expert systems and the bubble will burst. But what happens when we solve those problems, when we achieve the additional breakthroughs that are required to produce general purpose AI, which would be AI systems that are capable of quickly learning to do anything that human beings can do, and probably a lot more beside. And so the concerns I have there, one of them is the problem of control, so if you're making systems that are more powerful than human beings, how do human being retain power over those systems forever? And that almost sounds like an impossibility.

And then I think the other big set of questions has to do with how do we find purpose in such a world? So even if we do retain control, we are then faced with what Kane's called our real, our permanent problem, which is how to live wisely, agreeably and well in the absence of any forces that arise from economic necessity. And those forces have shaped our society since before we were human beings, and the economic would include the need to find food, the need to ensure safety of yourself and your family and your tribe, and shelter, and so on. This is what's been driving human life and has given people a reason to get out of bed in the morning for hundreds of thousands of years, and what happens when that goes away?

DSW: So Stuart, get for us a nightmare scenario, how could things go really badly wrong? It's interesting to have that, to juxtapose it to the optimistic conception of the Noosphere. Describe for us a really dystopic outcome.

SR: Films and novels are full of dystopian outcomes. In fact, the really hard thing is to describe a utopian outcome. And so in some sense, I'll begin by complimenting your request, which is to say we've run many workshops where we've had economists, futurists, AI researchers, science fiction writers come together and try to describe a utopian outcome with general purpose AI, which is able to do all of the things we currently think of as work. So just focusing on the socioeconomic arrangements, it's really hard for people to come up with a positive vision because you keep getting stuck on this fact that the economic forces will end up taking away all the economic roles from humans.

And are we able to find meaning in our lives in the absence of those roles? And I'm inherently an optimistic person, and I would like to think, yes, we can, but it's very hard for people to describe what that would be like. And I think part of it is because we're all trapped, we're all products of an existing system of education and acculturation, where our work is such an important part of who we are, and it's why we have an education system, is to produce people who are capable of fitting into all the economic roles that have to be filled.

So my view, and this is something that many previous thinkers have reached as well, is that we're going to need a completely different system of education, and one that's based on science that doesn't yet exist, which is really a much more humanistic science. If I had my cell phone with me, I would take it out and say, "Look at this thing." We have put at least a trillion dollars worth of R&D expenditures to get to

this thing, the cell phone, which may or may not be ruining everybody's life. And we've put a tiny fraction of that into understanding what makes human beings happy, what enables us to live fulfilling and productive and socially engaged lives, how one person can help another do all those things. We don't understand the beginnings of the answers to those questions, but those are the answers we're going to need to have in the future, and it's going to take a long time for us to get from here to there, and we may not finish in time, and so that does worry me.

TD: So I want to frame the problem as follows. So I mean, I'm a student of cognitive science and psychology, as well as neuroscience and evolution. And one of the histories of psychology, of course, that we passed through in the 20th century, was this massive move towards behaviorism. And I consider what we do with AI as a rediscovery in machines, what behaviorism was about, controlling animals and so on. And let me see if I can frame this for you, because I actually see both the vision that AI has produced as being the way that behaviorists looked at organisms and animals and humans. That is, all you had to worry about was input and output, and you construct a system that produced, given the right inputs of the right rewards and punishments, the right experiences, you got the outputs that you wanted, the goals that you wanted. But in that process, it allowed you to say, "Well, I don't have to worry about what's going on inside."

The other side of the behaviorists story was of course, that it produced this remarkable set of tools for controlling behavior, remarkable set of tools that said if we tweak this and that by this and that reinforcement, we can actually create all kinds of remarkable behaviors once we understand a little bit about the basic motivational system of the organism. In the case of behaviorism was oftentimes as simple as pleasure and pain, hunger, and thirst. But I see the other side of the AI story as the perfect behaviorist trainer of human beings. And why do I say this? And that's because a good behavioral system figured out how to use incremental changes in reinforcement to structure behavior in a way that might be below the consciousness of the individual organism or human being, being modified.

The thing about artificial intelligence that is to me, part of this dangerous look towards the future, or maybe even a positive look, depending on how you perceive it. And that is that it's capable of tracking minute details of human desires, human movements, human purchasing, comments that we leave online, and so on. Using that information, as we do crudely with recommender systems today, to tweak it slightly, to tweak the environment slightly, changing the reinforcement surround. With artificial intelligence, where you have the capacity to use billions of data points from a person's life and experience online, to do this, you have the perfect behaviorist trainer also. And it seems to me that one of the ways we can look at the future is both for good and ill. China, I think, is trying to do this in real time, with facial recognition and credit systems, to actually find a way to systematically, behavioristly, train a whole population. I do think that that's one of the futures of this system. We've learned how to do it already in a very small, but you might say, inappropriate way, just by accident, with Facebook and with Google and with Amazon, but we're learning how to do that, and as a means of manipulating and controlling, it seems to be an almost inevitable future.

DSW: All right. So I'm so eager to take my turn here. And just yesterday, I had a conversation with Kevin Kelly on this same topic, in this series, and his book, What Technology Wants. And in that book, he spends quite a lot of time on the Amish, which seems strange, but it turns out that the Amish are actually technologically quite savvy. And the thing about the Amish is when they decide whether or not to adopt a certain technology, there's always one question that's foremost on their minds, which is, what will be the impact of this technology on our community? What will be the impact on our community? That's the overriding factor as to whether they have electricity in the shop, but not the home, and whether they drive by horse and buggy or by car, and many other decisions, what will be the effect on the community?

And it's very easy to see in this case, that if you suspended that, and if you just allowed individuals to make their choices, as opposed to having this community selection criterion, community welfare criterion, then the community would fall apart. We find this very, very easy to see. But the next step is to generalize it and to show that's actually true in each and every case, that if we want a system to function well, we need to ask the question and we're evaluating options, any options, technological options or not, what's the effect going to be on the whole system? And then to select the options on that basis, on the basis of that systemic goal. And if you see that, I mean, that's all around us once you know what to look for, it's in software development, it's in just about any systemic goal. That's what people do, they try stuff out, and then they select them on the basis of the whole system.

But the end result of that thinking is that we have to do that ultimately for the community of the whole earth. And if we're selecting our options on the basis of lower level entities, Facebook being one, then we're going to get the equivalent of the Amish community, in which individuals are allowed to make their choices, and you're not filtering those choices on the basis of the whole community. So I think a lot of our problems are based on that. What it leads to, I think, inevitably, is a whole earth ethic, in which when we evaluate our options and when we train our AI systems, then that's what we need to train them to do. And if we do anything less than that, then we're going to get problems. So actually that's a pretty big piece all by itself. I have some more things to say, but I'm going to wait until my next turn, in order to get to those. So Stuart back to you, you can reflect upon anything you want, I think it's just fun to take rounds on this and just see where the conversation goes.

SR: I want to take some of what Terry said and disagree with it a little bit, and I think you can use behaviorism in two ways. I think a history of cognitive science and AI would probably say that in a sense, AI eliminated behaviorism as a way of thinking about organisms. Because behaviorism for bad, any internal model, whereas AI said, no, no internal models are just as valid as theoretical descriptions of behavior, as electrons are for models of atoms. And so AI models in the hands of people like Newell and Simon, involved goals and beliefs and deliberation and all kinds of internal things. So it made for a much richer picture and this scientific validity of a much richer picture of what was happening in the mind.

So that's one thing, but I think the other point that you're making Terry, is that the mindset of how we use AI is similar to the behavior mindset for controlling, in the sense that we specify some objective. And so in social media, for example, there's an objective of maximizing click through. And when you optimize that objective, the algorithms end up manipulating people, changing who they are. In fact, I think changing their preferences so that in future, they're more predictable in their response to items that you would like them to click on. And so the algorithms simply function to manipulate individuals towards whatever is the closest point where they're maximally predictable.

And so in that sense, the behaviorist tendencies in the AI community are, I think, really harmful. And this comes back to the control question and also David, to your question about dystopias. The dystopia that I'm most concerned about is AI systems being developed, having more and more influence over how things evolve in the world, and optimizing objectives that originally we put in like click through or eyeballs or engagement, but objectives that are not complete and correct pictures of our preferences about the future, but typically very narrow, very partial. And optimizing those objectives ends up wrecking everything else, and I think that's what we're seeing.

And so the dystopian picture is that the AI systems become more and more capable and it becomes harder and harder for us to prevent what we may even see as a slide into a dystopian future, because of the way the AI systems are controlling us. So you could think of what's happening with the fossil fuel industry as an early warning. Fossil fuel industry, I mean, it has human components, but it's basically a machine that's designed to optimize some complex combination of discounted profit and market share and maybe something else. And in the process of optimizing that objective, it's destroying the world, and yet there's nothing we can do about it. We have lost this competition.

DSW: On that point, before we get to Terry, can't we just say that about laissez faire capitalism, with or without technology?

SR: Yeah. If by laissez faire capitalism, you mean ignoring-

DSW: No, it's everyone pursues, greed is good, greed is good, everyone pursues their self interest, invisible hand.

SR: Yeah. But the invisible hand doesn't say you should ignore externalities. For that to happen, you also have to have regulatory capture on all the other things, and disinformation campaigns, which the fossil fuel industry began 60 years ago, long before the rest of us were really all that worried about climate change, they knew that it was going to happen, they knew it was going to crimp their freedom of action and possibly even put them out of business. So they made sure that we couldn't do anything about it, by capturing the political sphere and pumping out disinformation and so on. So they won, and they're just a simple machine made out of humans. So the dystopian concern is, imagine that on steroids, where we end up having very little understanding or control over what's happening to us.

And I keep having to emphasize this whenever I talk to journalists, it's not spooky emergent consciousness, it's not that the machines will decide they don't like us and want to destroy us, they're just carrying out the objectives that we put in. But that whole approach, the behaviorist approach, if you like, of saying, this is the objective, and we're just going to build machines that should optimize it, it just doesn't work. And you can see with individual human beings, I'm sure Terry, you know about this sordid history of what some people call wire heading, which is connecting a wire directly into your pleasure center and giving you control over the stimulation of that wire.

When you do this for rats, they simply kill themselves. They keep pressing the lever, they stop eating, they sit there in their own feces and urine, they don't clean themselves, they just keep pressing this button until they die. And human beings, the experiments didn't go that long, but all the signs, I mean, the literally people sat there in their own wastes, pressing this button for 24 hours, and all the signs suggest that we would kill ourselves if we had such a button. And this is another example, it's evolution built this objective into us, evolution was a little behaviorist, it had a behavioral theory of organisms.

DSW: And selection by consequences, that's what Skinner called it, selection by consequence.

SR: It set us up with this a proximate objective, right? It's the objective of maximizing dopamine production or whatever, doesn't exactly line up with evolutionary success. And it can be hijacked in some circumstances and lead to evolutionary failure and individual failure, and we should learn these lessons. And what I've been talking about in my book, Human Compatible, is actually getting rid of this way of thinking about AI altogether, the idea that we specify objectives, we put them into machines, the machines optimize the objectives. That's a mistake, it's a seductive mistake because it looks like you're getting technology to do your bidding, but the world's cultures are full of stories saying it's a mistake, it's King Midas, sorcerer's apprentice, everything with genies and jinns and all the other stories in all these cultures about the failure-

DSW: The wish that you might make and end up regretting.

SR: The third wish is undo the first two wishes, because I may ruin the world.

TD: My favorite version of that story is the golem story. The golem, I actually titled a chapter of one of my books, golem. The golem is of course created as a device, in this case, an animate human-like creature that has all power. And it was to defend, in one of the classic stories, to defend the Jews of Prague, this golem monster was created, but was given a task to defend, to save, it's again, the three wishes problem. Interestingly enough, my favorite version of it, you activate the golem, you make him out of clay and you activate him by writing the word truth on his forehead, which is emet, in the Yiddish

of the time. And it was given a set of commands that had a particular end that it was supposed to produce, and it would do so to the letter of the precise law.

And so you give it commands and it follows through, it's a truth preserving algorithm, so to speak, that does things in the world. The problem is of course, like anything else without discernment, it begins to do terrible things. And the only way to destroy it was to modify the word truth, emet, to turn it into met, which means death. And it's interesting, in this particular myth, truth preservation has death as its core, and I see that as a wonderful allegory to the situation of machines. Machines are determinate, they produce things, and if we do it right, they produce them inerrantly, and the problem is, of course, there's no discernment in that process. And one of the things that we're asking here is how do we get discernment into the golem, and one in which we can say, "Okay, wait a minute, that's too far."

And I think Stuart's issues about control, and this issue that you bring up David, about the Amish, always being able to say, "No, wait a minute, before we do this, let's get a sense of it." The problem with all of this is that we don't yet have the predictive power to do this. In fact, we try to do simulations to figure these things out, even with climate change, now we're desperately trying to come up with simulations to figure out where we go, and if we get pushed to this moment of geoengineering to try to stop it, the simulations will even become more important, almost certainly there'll be golem effects there as well. My point in all of this, and I think it goes back to Stuart's analogy to magnetization. These are self-organizing processes, these are processes that amplify little bits of things, very, very intensely and very rapidly.

So little biases in recommender systems, if just let run billions and billions of times, really biases a system. We've got now an information biasing system running the world, that has this capacity to do these runaway self organizing effects that are very hard to predict. And I think that's one of our big challenges, where it's one thing to say, "Now, is this a good thing or a bad thing? Let's see where it goes." We can't even do that prediction. And in fact, we need good AI to do the predictions for what AI might do, which is, I think, part of the circular challenge that we're facing in all of this. But I think the real issue is that we need to spend a whole lot more time studying those kinds of processes and figuring out how AI itself can be used to pursue this.

My analogy to this is also these days, the shift from person hacking to AI based hacking. One of the things that's happening worldwide now, in terms of hacking information systems that have been set up to protect against hacking, is you've got to get a good AI system that can figure out what the protection system is and work around it. And so there's an AI hacking cold war developing, as there is, and Stuart has a lot to say about AI weaponry and the cold war, that potentially could be a hot war of AI weaponry as well. It's because AI is probably also the tool we need to anticipate the effects of AI, and that, I think, is a really interesting challenge. And it's not just asking what we want, but it's trying to anticipate what these consequences will be, and I think that's our big challenge.

DSW: Yeah. So in a lot of these conversations, I've emphasized a distinction between two meanings of complex adaptive systems. Meaning number one is a complex system that's adaptive as a system. Meaning number two is a complex system composed of agents following their respective adaptive strategies. And when it's the second complex system, then it's never going to work well at the whole system level. It's not the case that the CAS2 system, composed of agents following their respective adaptive strategies, just self-organize into a CAS1 system, a system that's adaptive as a system. A CAS2 system gets you your hacking AI, of course, it's oppositional, it's agents in conflict with each other, that will never function well as a system. And the more AI improves those competitive and disruptive skills, then the more disruptive the system will get.

But now here's something that's a little bit more optimistic, I think we can say about the golem problem and the Midas problem, and so on, which is this. The only solution to the problem of unforeseen consequences is humble experimentation. We need to make our best guesses, and then we need to try

it out, either in the form of simulation models or real world efforts, we need to assess their consequences with respect to what we're trying to maximize, and then we need to do that again and again and again. And if our target of selection is the whole system, so now we're working towards a CAS1 system that functions well as a system, and if we take a humble experimental approach, well, there's our best chance to achieve our goals. And the better our AI systems get, then the better they'll be at doing that.

So I actually think that that is an optimistic prognosis, and in that case, when you think of using AI to help experiment, and of course, so many AI algorithms are evolutionary algorithms, they're doing variation selection algorithms, like on warp drive. So those are the simulations that are basically getting you the candidate solutions. If you combine those with actually real world interventions, then AI is not operating on its own, like some kind of autonomous system, and there's humans that they're stuck there doing nothing. That's not the case. It's much more of an integrated interaction between the human effort, so now humans are working, you don't have the problem of just humans having no work to do. I think that's something pretty wrong headed about that. There's huge things to be done and humans are using AI as a tool to do it. So is there anything wrong with that picture Stuart? In some ways I've been channeling Tim O'Reilly on this topic, he's another person that I think is a good thinker on this topic and will be included in this series.

SR: Well, you said a lot of things and probably some of them are right, and some of them wrong. So I absolutely agree that experimentation and probably in simulation would be best, just as with geoengineering, you don't want to be experimenting with the whole planet seeing if that was a good idea or not. And I would argue that yes, we should probably build experimental facilities, sort of Sim City, except with a lot more depth and variation and so on. But I also really think there's a role for theoretical understanding.

And obviously it's a cliche that you have to have theoretical understanding to even drive the things that you try in the experimental setting. But the proposals that we've been working on for provably beneficial AI, involve a different type of AI system, an AI system that knows that it doesn't know what the human preferences are that it's supposed to be satisfying. And you can set this up as a mathematical problem, where the human is the one that has preferences, the machine, it's the same payoff, the payoff for the machine is the satisfaction of human preferences, but the machine doesn't know what those preferences are. And then this creates a game, in the game theoretic sense of a multi agent decision problem, and then you can solve those games. You can actually analyze the properties of the solutions and look at what happens.

And indeed you can show that at least under suitable assumptions in the limit, the solutions are beneficial to the humans, no matter what preferences they have, the behavior of the machine ends up being beneficial to the humans. And that's not the case in the classical model of AI, because if a machine has fixed objectives and they aren't already perfectly aligned with the humans, then the outcome is arbitrarily bad for the human, because the machine ends up setting all the other variables to plus or minus infinity. You get what the control theory is called, bang, bang control, setting things to extreme values because that enables you to get a little bit more juice on the objective, and that's almost always disastrous.

So there is this role for theory, and the places where we don't yet have good theory, so we can look at one human, one machine systems, we can start to get some progress on many human one machine systems, it gets really hard with many humans many machines, when those machines are in some sense, uncoordinated. So imagine each one is produced by a different corporation, they all conform to the general provably beneficial AI template, but we don't yet know how to get them all to avoid prisoners dilemmas and all the other things that happen in these systems.

DSW: You might need something like moral norms in order to do that, just the way with humans.

SR: Yeah, more than that.

DSW: Go ahead. Go ahead, Stuart.

SR: Yeah, I mean, in some sense they have moral norms, their goal is precisely the benefit of humanity. But even given that, you could still have prisoners dilemma situation emerging, where they all want to optimize for the benefit of humanity, but they end up with collective actions that are actually harmful to humanity. And so we need to get a better understanding of how to get what some people call zero shot coordination, but that they're dumped in the world and they figure out how to cooperate with each other to avoid these prisoners dilemma solutions.

TD: I wanted to just go to a more practical issue for a second, and that is, we just came out of a large scale international meeting about the climate and what to do about it. And everybody seems to know what the bad news is and what needs to be done, but of course at the level of political decision making, we basically, after 20 some of these meetings, still haven't got to the point where we can actually say we can implement this solution. I've become relatively cynical about political top down solutions, whether getting corporations to work together, nations to work together, different religious groups to work together and so on.

And this is one of the real challenges, is that the adversarial use in some way or other of AI, looks like it's almost more difficult to interrupt than even these sorts of things, in part because once we set it running, it's such an amplification system that it's hard to get a grip on it. We've done that, obviously our climate change problem is an amplification problem, that is, we can't stop producing carbon in the atmosphere because we're so addicted to what we get from it, the plastic, the mobility, the warmth in the winter and so on and so forth. The addiction has forced us into this and we're finding it's very difficult to back out of it.

So part of what I'm wondering is yes, although we might even run great simulations and figure out exactly what we need, in many respects, the human context of this tends to be working against it. And I'm wondering if there is a logic to the AI system, that is the fact that we've now got a global communication system set up in which AI could play a role in determining who talks to whom and what information gets passed on and what doesn't. Is that also just a recipe for a user problem, a control problem, or is there actually another way around it, that is a way around the political impasses.

SR: I think at the moment the AI systems that are controlling a lot of the communication connections that happen in the world, they don't even know that human beings exist. So the possibility that they could think about and solve these multi level, multi stakeholder coordination problems, that's not really on the cards as yet. But I'm cautiously optimistic about the possibility of cooperating on the control problem, because it's immediately obvious to everybody that we do not want to lose control. And if there's anyone who doesn't want to lose control, it's heads of state. So it's somewhat reassuring that Xi Jinping is talking about the existential threat that AI poses to mankind.

TD: And yet he's one of the big users these days.

SR: Yes. I mean, but the kind of AI that China needs to do, what China wants to do with its social control policies and security and facial recognition, all the rest, is much more limited. And so I think that if we're able to come up with viable solutions to the control problem, there'll be actually a political incentive for everyone to adopt them, as well as an economic incentive. But these will be systems that actually work much better for users, they will avoid cooking the cat for dinner because they will either learn about human preferences for their pets, or they will say, "Well, I don't know about human preferences for their pets, but they might be quite significant one way or the other, and so I will hold off or ask permission before-"

TD: Would you call this a kind of humble AI?

SR: Yeah, exactly, I've used exactly that phrase, this humble AI, and there's all kinds of ways of contrasting it to existing AI systems, but knowing that you don't know what the objective is, is a form of humility, and it leads you to actually interact with the world in a minimally invasive way. You don't want to change things unless you know for sure that humans want them changed in that direction.

DSW: I wanted to bring up, pick up the behaviorism theme and a report on the current generation of behaviorism. As we know historically, behaviorism was supplement by cognitive therapy, behavior therapy was supplemented by cognitive therapy, and then cognitive therapy was supplemented by something called mindfulness based therapy. And so the most effective therapies today are called behavior cognitive and mindfulness based therapies. What does that mean? It means, among other things, that first of all, centrally important symbolic thought. That this is not just like responding, operant conditioning, trial and error operant conditioning, every person has a rich symbolic system that's driving their behavior. And so therapy basically involves working with the symbolic system of a person and it needs to be value based. First of all, you need to clarify, there are individuals in groups, what are your values, and then you have to appreciate there's other selective forces that are driving you away from values. That's why we find it hard to keep our new year's resolutions, we've got our aspirations, typically we don't keep them.

That's true at the group level, in addition to the individual level. And so therefore you have to actually face up to the fact that there are selection pressures operating in your life that are taking you away from your valued goals, and so you have to accept them and commit to working around them. So I've just described something called acceptance and commitment training. Isn't it interesting, I think, that if we're really got to talk about behaviorism, let's talk about the current generation, and if we do, then I think we've really got some solutions that we can think about. These ideas have not been applied to AI to my knowledge, but I think that they should, and I think I'd be very happy to help lead an effort along those lines.

(short break)

DSW: Is there an interesting comparison analog between the autonomic and sympathetic nervous system, something that's relatively automated, as opposed to something that's more reflective? We know that our nervous system is specialized in that way. Can we say something similar about AI in the Noosphere?

TD: Let me start with the neuroscience of it, and that is the autonomic and the parasympathetic, sympathetic nervous system, which run digestion, breathing, heart rate, basic metabolic functions. These are things that are run in background, we're not conscious of them, we feel them when they're out of sync. In the evolution of nervous systems, there are nervous systems that were basically just that. So if you look in particular organisms that are radially symmetric, they mostly don't have heads and tails and eye spots and mouths and so on. And as a result, their nervous system is quite distributed and things are locally maintained. When animals developed, they had a head end in which most of the work was done, in which most of the prediction, because once you start moving around, you have to predict the future.

You have to have organs that can see into the distance, can hear in the distance and so on, or not just immediate. So that oftentimes the nervous system in animals is broken up into this sort of, you might say the predictive nervous system and the maintenance nervous system, they're linked together for obvious reasons. One of the questions is right now, I think that largely we're dealing with this linkage of AI into human interactions, as a more autonomic maintaining, runs the stock market these days, and things like that, background stuff, keeping track of traffic and that sort of thing.

The question is, are we more like a starfish with a distributed nervous system, or are we developing a head? A head end is one that predicts the future, and the two have to be the linked together in a

complicated way. In the course of evolution, it was the metabolic system that came first and it became more and more sophisticated, and only when it began to have to deal with prediction, did heads develop, did brains develop. My guess is that the state we're in now is still autonomic, but it's moving away from that. And this is the Noosphere question, probably carried out this analogy to a global brain, I think the analogy is too simple for all of these reasons, even in terms of talking about what a brain is, brains are very different in different species for different reasons.

But one of the questions I'm trying to pursue is, and I think this is hidden in Ben's interests as well, to what extent will we be able to rely on AI to do these autonomic functions, so that we don't have to pay attention to them? Run our countries, run our energy systems, run our fuel systems, our food systems, our distributing systems, and so on, in a way that we don't have to think about it, it's just handed off. And we have to have then this other part of the system, which is the look ahead, the experiment, the trial and the error that we've been talking about, that's what brains are mostly about.

DSW: Oh, that's great. But there are forecasting systems, so I think they both exist to a degree. I think that there are the forward looking predictive algorithms, even before technology, there were forecasting decision making processes and so on, scenario planning and that kind of thing. So it does exist to degree, but I love the distinction, Terry, I love the way you described it because it juxtaposes, especially when we get to Stuart, basically the biological realm and now the human technological realm and how they're so similar to each other. So Stuart, what do you have to say about this?

SR: So I think that you could ask all these questions in the complete absence of technology. You could look at the whole human social and economic system and ask, how do we draw analogies to the autonomic nervous system or the sympathetic nervous system. And there are lots of parts of a system that do sort of work like the metabolic system, they just function in the background. When you think about the system of food production and distribution, it's incredibly complex and yet incredibly effective that somehow we manage to deliver 20 odd billion meals a day to seven odd billion people.

And it does break down a little bit here and there, but overall, despite its huge complexity, it functions pretty well, and a lot of that has to do with the way markets operate. So to a large extent, we have turned over a lot of the management of our human system, to markets, which are in a sense algorithms, and interestingly, we did a little bit of work in the 18th century onwards, to try to reassure ourselves that they would actually function, but they were functioning for thousands of years before that, with no theory, they just emerged and they seem to work well enough to fulfill those autonomic roles. And I think we haven't done the due diligence for what happens when we allow AI systems, first of all, to function as participants in those markets, and then even to replace the market mechanism altogether with some new kind of decision making or control system.

And I think with a flash crash, we saw that, oh yeah, it doesn't work, when you replace enough of the human actors in the market, with AI actors, you can get these, just like with the magnetic domains, you get this failure because of systems being too similar to each other, or somehow causing each other to operate in sync, in ways that human beings don't, and you've got these massive failures occurring. And if we look at where else are we putting AI systems in charge of these autonomic functions, employment markets are increasingly, both in telling people where to apply, and on the corporate side, deciding who's going to get employed.

That's a huge function that's been working moderately successfully, although different countries have different systems. Some countries actually disallow employment agencies, that's an illegal function in some countries, but not in others. And so there are different designs, but putting AI in charge of that, we've already seen with Amazon that, oh, they didn't even realize that their algorithm was rejecting all female applicants out of hand for technical positions. Well, not all female, but if your resume mentioned the word women's, like I was a member of a women's choir or a women's rugby team, or I support this

organization for women in technology or something, it would just reject your application and they didn't even realize that was going on.

DSW: Was that a feature or a bug perhaps? Was that a feature or a bug? Is that intentional-

SR: No, it was completely unintentional, it was just a result of correlations that the algorithms noticed in the previous decisions that were made by Amazon's hiring managers. I think it over emphasized there and ended up being a disaster, both in public relations terms and in the effect on applicants and Amazon's business. So I think again, you see this problem that you move from a large number of heterogeneous agents, to the possibility of these coincidental alignments, where the things start to happen in synchrony and the stability of the system goes out the window. And I think other areas, a lot of eCommerce now, particularly business to business, is becoming more and more automated in terms of selection of vendors and agreement of prices, and so on, is moving towards full automation.

And again, we could start to see large sectors of the economy where it's actually AI algorithms that are deciding how that sector evolves, its structure, the flows of money and goods, and so on. And just like the employment market, that could go wrong in ways that we might regret. So I think we have to be ... And I think social networks are another thing, to a large extent, AI algorithms are telling people who they should talk to and who they should get to know and who they should mate with, and who knows what effect that could have. We used to have human matchmakers and they had certain defects as well, but we've relied on systems that involve lots of heterogeneity, lots of serendipity, and so on.

And I use this analogy actually for talking about whether we should edit the genes of our offspring, but I think it applies here too, which is that we're in this casino, we're all playing roulette and then someone just comes along and picks up the ball and puts it in the zero, and the element of chance is now gone and they collect all the money and leave. And is that a good thing, do you want to get rid of the element of chance and the systems that are built up around it, to actually benefit from all the chance interactions of all these heterogeneous elements. Our society's partly organized around that and to benefit from it, and if you take that away, it might be the way our society is organized, it no longer gets to benefit from that, and is not well adapted to this new kind of circumstance.

DSW: So I think that this has been a great conversation friends, and for me, it just highlights the fact that the Noosphere, the idea of some global thinking envelope, is not going to self-organize. I mean, it's extremely challenging to actually accomplish something like that. So we can't be sanguine about the concept of the Noosphere, it's something that we have to construct and it'll be extraordinarily difficult to do so, it's going to take the very, very best of our knowledge to do so. So I think it's a bit of a humbling conclusion to this conversation, but a very important one, because that's just the way it is. And I think that I really thank you for contributing your insights to that realistic conception of the Noosphere.

TD: If possible, I want to ask one last question before we run out of time, which we're about to do. And this has to do with what I think is underneath and behind all of this, and that is, is anybody home? That is, we're dealing with, when we think about thought and cognition and human experience, we're dealing with feeling, we're dealing with the possibility to suffer, for example, or to feel joy. One of the questions that one of my friends overseas, a man named Thomas Metzinger, for example, is advocating for a moratorium on conscious AI, by what he means is sentient AI, AI that has skin in the game, so to speak. And he said that yes, we're studying this in human beings, studying this in brains, should we hold off and not do this, not even try to experiment with this, have a moratorium on this in terms of our devices?

Now, the reason I bring this up, especially in the context of the autonomic versus the predictive nervous system issues, is the autonomic nervous system is a feeling nervous system, but it doesn't give that information forward until it needs to, until there's a problem. And the intelligence, when we talk about AI, we're talking about intelligence that doesn't feel, to some extent, that there's nobody home, it's solving problems, it's really great at this. But one of the challenges is we're turning more and more of

our lives over to an autonomic nervous system, to a system where there's nobody home, where there's not a somebody. And of course, many of the science fiction stories assume that the AI is going to want things and hate things and desire things and suffer and so on, it's going to have sentience.

But as far as we know, so far, brains are very different than this because brains evolved from creatures that had to keep themselves going, so to speak, at every moment they're about feeling their own existence. AI is not worried about its existence yet, that doesn't mean we can't do it. But I'm struggling here with, I think one of the worries we have in the back of our minds is on either side of that coin, both of them sound scary. That is the AI that has desires and fears and angers is scary, but also the AI that's just a machine is scary. And I think that neither of them feel comfortable, and I think this is one of the discomforts about the future.

SR: Well, I would agree. I think it's very hard to do anything about Dr. Metzinger's suggestion, for all we know we've already created conscious machines, we have absolutely no way of detecting consciousness in machines, or actually-

TD: We don't know what it is.

SR: Right. And this is, of course, a very old point, but we can't do it in humans, and as Turing said, there's a polite convention that other people think and have conscious experience, but it's just a polite convention. And so as far as I can see, there's nothing we can do about the concern. To me, the concern is not that the AI system would do anything different because it's conscious, it's still running C++, whether or not it's conscious. And so you can predict what it's going to do from analyzing the C++ and the consciousness doesn't give it any causal powers beyond that.

So the concern, the extra concern is that it then maybe it should have moral rights. And it goes from zero weight in our calculations of the future, to having a weight that might actually weigh against our own interests in the future, and that's something we should be concerned about. But I suspect what will happen is that we will simply blithely continue to be in blissful ignorance of whether our machines are conscious, and it might be that as they behave in more and more intelligent ways, that we will just start to assume, as we do for each other, that they are conscious, and again, we'll have no other form of empirical evidence that consciousness is happening.

DSW: I think what'll happen is that they begin to behave more and more like humans and we'll attribute consciousness to them, just the way we attribute consciousness to other humans. And then we'll have feelings towards them, we'll be loyal, we'll love them because of the way they respond to us. And so we will endow them with consciousness, the way we endow each other with consciousness.

SR: So this is something that we can regulate, we can say, that's just a bad idea, don't build machines that engender those kinds of responses in humans. So in particular, do not make them humanoid in form because that can bypass a whole lot of rational analysis.

TD: In fact, it's human gullibility that's the issue, and I think of the Turing test as a human gullibility test, that is, we build a machine that can now fool us. I think that turns out probably to be easier than we think.

DSW: All right. Okay guys. So I think this has been a wonderful conversation, one more time, and so thank you so much.