## Episode 4: Steve Omohundro - The ethical implications of artificial intelligence

**SO:** I've always been interested in the structure of ideas. That's sort of my real passion. When I was in high school, I really thought electronic music was really interesting and so I started a company to build electronic music synthesizers and thought that was what I would do. When I actually ended up in college, I found that I already knew a lot about the technology and I didn't feel like I would learn a lot, whereas physics and mathematics was really challenging. The questions I was really interested in as a young man were basically philosophy questions. What is the nature of ideas? What is the nature of meaning? Where does ethics arise from?

I had a fantasy back then of if we could only combine the questions of philosophy with the techniques of mathematics, we would have a very powerful tool for improving the world basically. I think that was my motivation. There wasn't really a discipline at the time that did that and really AI, that's in some sense what it is. It's bringing to bear powerful mathematical tools to address fundamental questions of humanity.

**KR:** Steve Omohundro has been working on the cutting edge of Artificial Intelligence for over 30 years. Part of his focus has switched from how we can create powerful, intelligent systems – to how we can protect humanity from the unintended consequences of building and unleashing systems which are exponentially more intelligent than us. This is his story.

**SO:** Around 2000, I began to realise that some of the systems I was analysing, studying and building had the potential to have a big impact on society. In particular, we were building what I call self-improving intelligence where systems that understand their own structure and intentionally redesign themselves to perform better. The question was what do they turn into after multiple iterations of this self-improvement? So that led me on a journey of trying to understand the human impact of these advance systems and how could we make sure that they have a beneficial, positive impact.

**[Introductory Music]**

**KR:** Welcome to How Researchers Changed the World. A podcast series supported by Taylor & Francis, which will demonstrate the real-world relevance, value and impact of academic research; and highlight the people and stories behind the research.

My name is Dr. Kaitlyn Regehr, I'm an academic researcher; an author and a scholar of digital and modern culture – and interested in how new technologies can broaden the reach and real-world impact of academic research.

In today's episode we are honoured to have been speaking with Steve Omohundro, a veteran of AI research and development, who's career spanning almost 4 decades includes: Physics, Mathematics, Machine Learning and, the topic of today's episode, Artificial Intelligence.

Steve works on a broad range of areas within the field of artificial intelligence. Today we will be focussing on one of the most topical: the social and ethical implications of artificial intelligence – in Steve's 2013 paper: Autonomous technology and the greater human good.

**SO:** My name is Steve Omohundro and I have been doing research in artificial intelligence for about 30 years now. As an undergraduate I studied physics and mathematics. Then I went to graduate school in physics, but I also had a mathematics advisor. I got my PhD in physics.

I was an undergrad at Stanford, they didn't even have a computer science major at the time, much less an AI major, but they did have classes. I took courses with John McCarthy, who was one of the fathers of AI. In fact, I think he was the one who invented the name artificial intelligence. That course was very influential on me. Doug Hofstadter was writing his book Gödel, Escher, Bach at the time and he had a course on that. I met him and had many discussions with him and that was very influential. I was sort of primed for AI work, but I ended up going into physics because I figured I would learn more in that field. I went to Berkeley and my roommate in Berkeley, Peter Bleeker, he was a student in AI at Stanford.

He and I would have lots of discussions on AI topics. In particular the work on machine vision, which was his focus. That gave me some published work which then helped my entry into that area once I finished with the physics PhD .

**KR:** The area which Steve was referring to, was of course Artificial Intelligence. Not the study of AI or even theorising – but building intelligent systems. This was back when many people didn't have a computer in their homes.

**SO:** When I finished the academia, a friend of mine was starting an artificial intelligence company called Thinking Machines and he recruited me. I went there and it was a very stimulating, amazing environment with advisors like Richard Feynman, so I spent a summer with him.

**KR:** Richard Feynman was an American theoretical physicist, known for his work in quantum mechanics, quantum electrodynamics, as well as particle physics for which he proposed the parton model. For his contributions to the development of quantum electrodynamics, Feynman, received the Nobel Prize in Physics in 1965. Feynman has been credited with pioneering the field of quantum computing and introducing the concept of nanotechnology.

Before we go on, let's find out from Steve, how he would define intelligence and artificial intelligence

**SO:** I define intelligence as the ability to choose actions that are likely to achieve goals. I would assign intelligence to, certainly the humans; we are very clearly intelligent, but even to animals like dogs; they certainly choose to do things to achieve their goals, and all the way down to single cell organisms and viruses. Then up towards the other direction, organisations have goals and they choose actions for that.  I believe all of those are intelligence systems, some more intelligent than others. You're more intelligent if you are better able to achieve your goals by choosing the right actions.

What artificial intelligence is, is trying to build computer systems that can do that. That they can sense what's happening in their environment. They have some goal that they're given to achieve, and they try and take actions to achieve that. For example, a chess playing machine. Its environment is the chess board. It looks at the current position. It has either implicitly or explicitly built into it the goal of winning the game of chess and it wants to choose moves that do that. If it's a good chess player, it'll choose good moves. It turned out it was pretty easy to write a good chess player because its world is quite simple, the chess board.

**KR:** So, with a relatively small amount of processing power, and the right minds behind the problem, computer scientists were able to create an AI system that could beat a chess champion. This happened in May 1997, when a computer developed by IBM known as 'Deep Blue' defeated Grandmaster Garry Kasparov – whom some would say is the greatest chess player to have ever lived. However, Development for Deep Blue began in 1985 – as a research project at Carnegie Mellon University.

But as Steve reminds us – the environment that AI system needed to master was only one chess board. How have things evolved in terms of capacity and capability?

**SO:** The question is as systems become more and more powerful; how will they behave in the world? One of the insights I had was one of the actions that any sufficiently powerful system can do, is to change itself. It could re-write its code. If it became sufficiently powerful and intelligent it could redesign its hardware, make better processing chips, maybe big, better robot arms. Those kinds of changes, changes to the system itself, impact its behaviour in its entire future. So, for many goals, by changing yourself you can impact that goal in a big way.

Any main goal will give rise to a bunch of sub-goals of things that would help it achieve that main goal.

**KR:** Now we aren't looking to tell the story of a sci-fi movie franchise, because sensationalism only serves to deny us the benefits – of technological advancements. So, we'll leave this to Hollywood. But there are genuine concerns that should influence future considerations in the field of AI.

**SO:** One of the things I realised is that for almost any simple goal, there are a number of sub-goals that appear regardless of what the main goal is. I called these the basic drives because they are kind of like the drives that animals have. One is that if the machine is turned off, for any simple goal like play good chess, if the machine is turned off it's not playing good chess so that goes against its goal. It will create a sub-goal of keep yourself from being turned off. That might be surprising to the designer.

Similarly, for almost any goal if you have more resources, more compute power, more electricity, more data that helps you do your goal better. Almost every system will have a basic drive towards getting more resources. Similarly, improving your structure, getting better algorithms, better hardware, that also improves things. There are a number of these things which for a broad range of goals, systems will try and achieve. They are very analogous to people. In general people try and prevent themselves from being killed, that's a self-protected instinct. People also want to get more resources and the extreme of that we often call greed.

In the Human sphere, we have a balance between social good and individual good. We have a whole bunch of moral systems, legal systems, religious systems which have arisen to help balance the interest and the needs of the individual or the interests and the needs of the society. I believe that we need to extend those systems to manage these new AI's that are coming in. They have somewhat different requirements and goals and so we're going to have to account for the nature of these new kinds of creatures, these new intelligences that are coming into our society. There are a bunch of hard decisions we're going to have to make. Should these systems be allowed to vote? Should they be full citizens? Should they be viewed as servants? Should they be viewed as slaves? Are they just machines? All those kinds of core questions come right into what the future of human society looks like.  And I think that understanding these basic AI drives, these potentially unintentional consequences of the goals that they have, is critical to understanding how to integrate them into society in a way that's good for humans.


**KR :**The different requirements Steve is talking about are largely dictated by logic. The problem is this, even comprehending let alone calculating probable outcomes is beyond our brain's capabilities

**SO:** If you look at the development of human society, our brand of humans, Homo Sapiens arose about 250, 000 years ago and for many years we were in small bands of 150 people, something called Dunbar's number. That small of a group, 150 people, every person can know every other person and we have these inbuilt emotions which are the moral emotions or social emotions where if somebody wrongs you, you know who it was. You recognise their face, you're mad at them. We had language so you could gossip about them and the group, in the extreme case, would ostracise them or kill them. That provided tremendous pressure for people to behave in ways that support other people in the group and if they deviated from that there would be a lot of punishment.

Humans managed to be among the most cooperative of all organisms and that is one of the things that led us to spread across the earth, to thrive and to end up being billions and billions of people as opposed to a small number of people. About 10, 000 years ago we developed agriculture which let us concentrate our storage of food and to create these communities that were much larger than 150. Our brains really don't have the capacity to manage 10,000 people as opposed to 150 people and so the violence level, the death rate and murder rate appears to have been quite high at that time. Steve Pinker has a book called The Better Angels of Our Nature which estimates that violence was hundreds of times larger than it is today at that time.

We needed additional mechanisms for us to work together in a positive way and the notion of a Government, the notion of some kind of Police force, the notion of using money as a means of organising interactions between people, all of those emerged say between 10,000 and 5,000 years ago and allowed human society to really flourish on a big scale.

Those mechanisms are based on the participants being human and these new systems that are coming, on the one hand they will probably be smarter than us, certainly in technical areas like breaking into a computer, that kind of thing they will be very good. Yet, their interests and their needs will be somewhat different. For a human if you die, that's it. That's the ultimate terrible thing and so Capital Punishment makes a big difference. For an AI system, if there's a back up copy of it, you shut of one copy or erase one copy, that's no big deal to it. So, there's a different set of requirements there.

**KR:** The point that Steve is making here is that humans, by our own frailty, have evolved to be cooperative. When these self-regulatory control measures didn't work, we were created our own enforced control measure – and call it society. The issue is, AI hasn't evolved according to the same principles as biological organisms.

**SO:** There's a big scientific question, can you design a societal structure, a set of laws and a set of enforcement mechanisms that apply to these agents, which potentially may be smarter than us? One of the challenges is that our current system is not set up for that. If you just plop down a super intelligent agent in the middle of today's society, if it weren't very carefully designed it could wreak havoc and there basically wouldn't be a good way of stopping it.

The intellectual question is, how do we get from where we are today to where we want to be, which is a society which uses the power of these intelligence systems, without them causing harm and negative problems?

**KR:** After the break, we'll hear how Steve wrestled with this problem and sought to discover a solution which might regulate and enforce safety measures on super-intelligent AI systems.

How can I make sure my published research has an impact on the world? What do I do if I disagree with peer reviewer comments? What techniques can I employ to manage my time in between teaching and research commitments? These are questions we hear all the time from researchers at different stages of their careers. We want to help. So, with the support of our partners Taylor & Francis we've created two 12-week learning programs to support your research career. For early career reseearchers we've covered the full process of publishing your research – from choosing a journal, to managing the review process, to boosting your personal profile after publication, and everything in between.

For mid-career researchers, our programme builds on your existing knowledge and experience of publishing your research to make the process more efficient for you. We've also included plenty of advice on raising the impact of your research, including driving discoverability using keywords and how to work effectively with journalists. If you think this would benefit you, sign up to a learning programme today at howresearchers.com/learning-programs.

We'd also like to take this opportunity to thank our supporting partners Taylor and Francis Group. We've worked with the team at Taylor & Francis Group to develop these learning programmes, drawing on over 200 years of their experience as one of the world's leading publishers of scholarly journals. It's thanks to them that these learning programmes are as comprehensive as they are.

**KR:** Before the break we were hearing from Steve Omohundro and his quest to develop a control system to guard against the unintended consequences of super-intelligent Ai.

**SO:** How do we get from where we are today to where we want to be, which is a society which uses the power of these intelligence systems, without them causing harm and negative problems? I had the idea of doing it in an incremental way, using one of the most powerful tools I believe is mathematical proof. With mathematical proof you can build systems where you have absolute mathematical guarantees about how they are going to behave. The trouble is our capacity for doing that today is pretty limited. Once we have more powerful systems, we can apply them to the task of creating systems which can be mathematically proved to be correct according to certain criteria.

We have the opportunity of using AI itself to develop more powerful AI and more governable and restricted AI. That gives the idea of having a structure, where at each stage we build something incrementally better that what we have today. We have very high confidence in its safety and in its value for creating a beneficial world. I call that sequence the scaffolding, sort of like the way ancient stone masons would build an arch. They would first build a wooden structure and then they'd build the arch on the wooden structure and then they could take away the wooden structure.

I think we need to do that in the AI world. We need to build a structure which gives us very high confidence of the positive behaviour of the systems. Once we've gotten up, use that scaffolding structure, we can end up with very powerful systems which will not behave in unexpected ways.

**KR:** So, by 'bridge building' or scaffolding we'd effectively be acknowledging and accounting for human fallibility?

**SO:** I think that strategy would work from a technological point of view. Where it's a little iffy is, we're also in a political environment right now and every country on the planet is developing AI. China has committed to

being the leader in AI by 2025. Vladimir Putin in Russia has said that the country which develops AI first will be the leader of the world. The United States certainly has a lot of AI. We're in a kind of arms race competitive environment so simultaneously doing things carefully and very precisely, with mathematical proofs of behaviour and battling for military AI with other countries. Those two don't quite go along with one another so whether that scaffolding vision is actually possible to implement is not clear to me at the moment.

**KR:** As well as the development of this powerful tool being akin to an arms race, what are the other ethical concerns we should be considering around AI?

**SO:** There is an increasing interest in ethics. Most of the discussion today is about issues which are very important but aren't really the central issues. The issues that are hot today, there's a lot of discussion and concern about privacy. Many of the social media companies, Google and Facebook especially, get a lot of press on this, make use of tracking your data as you use their services and they use that to target advertising to you. That makes their ads worth a lot more and that makes them make a lot of money. Early on I think people viewed that as a fair trade. You give me free use of a search engine, free use of an editor and all that, and you will serve me even better ads, I like that. I would say four or five years ago, there started to be a lot more concern about your personal data somehow being misused.

One of the challenges there is that political statements or news or advertising, can be targeted to your particular personality. The famous case there was Cambridge Analytica was using quite a simple AI to build personality models of people on Facebook based on what they had liked, and then used those personality models to target political advertising. Then they claimed to have helped sway the US Presidential election and I think even the Brexit election, elections around the world using that technology. Whether they did or not is still, I think, somewhat controversial but the concept that you could do that suddenly hit the main stream. People took a knee jerk reaction which is, "Oh my God, you can build models that will manipulate me from my data, you should not have access to my data."

One approach is to impose strong privacy on anything that individuals use. Apple is now using that as part of their brand – 'we're not going to use your data for anything'. I think going forward, that one is going to have to soften. We're going to have to find a different sort of perspective. First of all, we leave data everywhere. We're lumbering creatures and there's going to be cameras everywhere, so we are going to have to live with the fact that most of what we do is not going to be private. It's more like, what happens to this and how do we deal with the possibility of being manipulated and lied to.

Another advance, maybe about a year ago, was that these deep learning systems were able to...there was a system called Deepfakes, where somebody was able to take an image or a few images of your face and a video of someone else doing something and recreate the video with you face in there.

And so, that created a shock. Both because you could make a video which looks like you're doing something that you never actually did, and you could make very realistic looking videos which could serve to promote false narratives. Fake news, that kind of thing.

**KR:** You may have seen the deep fakes video of Obama being pupated by Jordan Peel. The video was created to demonstrate how AI could be misused for political reasons.

**SO:** So, you can make Obama say anything and even more scary, you could make a video of Trump saying, "We're bombing North Korea tomorrow." Which, if the North Koreans believe that, might cause them to act and you could potentially stimulate a nuclear exchange based on a fake video. That's a huge worry.

So, one ethical issue is privacy, using your data. Another one is creating fake artefacts which convince you that something is true which actually isn't. A variant of that is manipulating you based on knowledge of you and what you like and dislike, convincing you to do things that maybe aren't in your own interest. Keeping a ground truth, I believe we need a truth machine. We need an infrastructure which helps society know what actually happened and what didn't.

There is some potential, I've written a bunch of papers and things on using blockchain technology to validate that a particular image, video or audio was recorded at a certain time, at a certain place and hasn't been modified. If we had that as an infrastructure, I think that would help us be better at creating consensus truth for humanity.

It's a battle. It's an arms race between using AI technologies to fake stuff and using other kinds of cryptographic technologies to have more confidence in it. AI can also be used to detect these fakes so it's kind of an arms race situation there. Certainly, that is an ethical question.

**KR:** So, we've discussed a bunch of concerns and ethical dilemmas – but this wouldn't be a balanced discussion if we didn't talk about some of the incredible, exciting and hopeful possibilities AI can bring.

Thankfully, brilliant minds like Steve's are not only balancing the ethical considerations, but working on solutions to real-world problems.

**SO:** I am now Chief Scientist at a company called AIBrain. Our focus there is on something we call augmented intelligence which is instead of...it's easy to slip into a view of the AI systems versus the humans. That AI's are going to take over our jobs, AI's are going to do this and what are the humans going to do. Our view is that no, we should view them as agents that support us. We're building these systems; we should be building them for our own benefit. Humans have knowns irrationalities and knowns issues and knowns problems. Why not use these AI technologies to help us be our best? To help us learn better. To help us deal with Psychological issues. To help us be more effective in our relationships, how to communicate better, a better emotional intelligence.

I think there is tremendous opportunities to use these very same systems for the growth and benefit of individual humans. Then ultimately for the harmonious interactions of humans creating worldwide peace, worldwide prosperity. If we make the right choices over the next decade or two, I think we have the potential to really create a utopia. I'm really actually very optimistic about the positive possibilities.

There's a group that identified the 15 biggest problems in the world. Things like global warming, economic inequality. We went through systematically and every single one of those 15 can be positively impacted by using AI and robotics. I think almost all of today's problems, assuming we can avoid the new problems; almost all of today's problems can be dramatically improved by AI.

**KR:** We wanted to ask Steve what he sees in our immediate future with regards to AI. Just as the dawn of the agricultural and industrial revolutions saw periods of turmoil; sociological and environmental – what will the dawn of the AI revolution look like – and when will that happen?

**SO:** I think we're in it right now. This is the moment. I see three phases coming. I think right now I call them the AI economy, the AI military and the AI society. I think right now we're in the AI economy, which is basically massive investment. The consulting firm McKinsey, estimated that just today's AI, using deep learning systems for recognition, chatbots, translation, those kinds of things, that by 2030, $70,000,000,000,000 of value will be created. Just for comparison the United States GDP is $18,000,000,000,000. We have a massive tsunami of economic value coming just from the low hanging fruit in the economy today.

I think we are in that phase where lots of start ups are flooding the deep learning classes. Stanford which is right near where I live, they have a few classes deep learning in natural language and deep learning in general and they are packed to the gills. Every student wants to take it. They're all doing these amazing projects and so there is just an excitement. I recently gave the keynote speak at a thing called Ling Hacks, which is high school students who are doing computational linguistics, in high school you know. There were 200 young kids, eager and enthusiastic to build their new chat system or their thing using language in some way. There is a sort of excitement and sense of possibility among the younger people today and that is just going to sweep through the economics of society.

At the same time, we are just starting to get some of the more sophisticated systems into the military. Military's advance on their own pace. We haven't had the first AI battle yet. Some people say that the South China sea, that the Chinese are putting autonomous robots on these fake islands that they're building, to protect them. It may be that the first robot-robot battle happens in the South China sea. That will be a water shed moment. That's going to make people think a lot when that happens.

The next phase that I call the AI society, which is some of the topics we've been talking about where we take the entire body of law and that becomes digital and we start having AI judges. We start having AI lawmakers or assistants to law makers who analyse the much more detailed implications of a proposed law. Who figure out how things interact with one another and eventually either AI politicians or AI assistants to politicians, could actually if done right, help prevent a lot of the corruption that we see in various countries today. That's where all these issues of what is a good set of laws and how do we govern these autonomous systems at the same time as humans and I would say for the benefit of the humans? To make sure that the world we're creating is one that we really want to live in. I think that's the third phase.

So, I think we have those three phases and they're in various phases at the moment but they're happening. I would guess it's a question of decades before those really become critically important.

**KR:** And how might we manage this period of global, political and social change?

**SO:** I think identifying what it is that are the core human values that we would like a future society to express and then creating this intelligent system architecture to support that. Rather than blindly building this stuff which may end up doing things that we don't like.

**KR:** To find out more about this podcast and today's topic, visit howresearchers.com/ai. In the next episode of How Researchers we're speaking with Dr. Girija Kaimal and unpacking her study into the Reduction of Cortisol Levels and Participant Responses Following Art Making.

We'd love to hear your feedback so please follow us on Twitter, Facebook or LinkedIn at @howresearchers

This podcast was written and produced by Monchu, recorded at Under the Apple Tree Studios. Our producers were Ryan Howe and Tabitha Whiting. Editing, mixing and mastering by Miles, Myerscough Harris at WBBC. We


would like to acknowledge the incredible support of Taylor and Francis group with a special thank you to Elaine Devine and Clare Dodd.

I'm Dr. Kaitlyn Regehr. Join us next time for 'How Researchers Changed the World.' Thanks for listening.