Is Artificial Intelligence Good or Evil?

DR. OREN ETZIONI

Chief Executive Officer

Allen Institute for Artificial Intelligence

The President's Distinguished Lecture Series • Stevens Institute of Technology

October 4, 2017

Thank you for the kind introduction. I'm really delighted and honored to be here. I'm going to talk quickly. This is not a technical talk. I want to set up the question and really get to question and answer and have discussion both at the end here, and I understand there's a reception afterwards. So, really, this is an invitation for a conversation. There's no right or wrong answer here, in my opinion.

Let me start by setting things up. Elon Musk, the tech entrepreneur, has told us that with AI we're summoning the demon, that AI represents an existential threat to the human race. We see headlines in the newspapers all the time. Here's one from *Newsweek*: "Artificial intelligence is coming and it could wipe us out." Is AI really poised to wipe us out? The famous roboticist, Rod Brooks, said if you're worried about the terminator, just close your door. Perhaps these robots are not as threatening as they're made out to be.

A video of Dr. Oren Etzioni's lecture, which includes his slide presentation, is available at **stevens.edu/lecture.**

Andrew Ng from Stanford, founder of Coursera, said that working to prevent AI from turning evil is like disrupting the space program to prevent overpopulation on Mars. It ignores the technical difficulties. It also ignores the potential benefits. Stephen Hawking, in recent remarks, said AI is either the best thing that will happen to the human race or the worst thing. It's nowhere in the middle. It's one or the other.

Before I continue, I'd love to get a sense from you. How many people are really worried about AI and its impact? How many people see it as a major positive for our society? Okay, it's a techy crowd. Maybe I should quit while I'm ahead. I don't know if you saw, a lot of hands went up on the positive side.

Let's talk about this in more depth. I think the first thing we need to do is really separate science from science fiction and the Hollywood hype from the realities of AI. We need to ask ourselves, 'What is AI today?' The answer to that question is certainly very different than what you see in Hollywood films, what you see in "Westworld," etc.

What is AI? First of all, for many years AI was a lot of hype. That's something that's changed. We have seen a number of very real AI successes. We've seen an AI program defeat the world champion in chess. We've seen AI do speech recognition very well. That's the basis of Siri, Alexa and systems like that. We have systems that can now do facial recognition with very high accuracy. That's used in security. That's used by Facebook. It's used in a variety of ways.

"We've seen an AI program defeat the world champion in chess. We've seen AI do speech recognition very well. That's the basis of Siri, Alexa and systems like that. We have systems that can now do facial recognition with very high accuracy."

We have very powerful machine translation, for those of you that have used translate.google.com or other such apps. Skype will simultaneously translate as you're talking to somebody. These are all major AI achievements over the last few years. Of course, we had IBM's Watson defeat the world champion in "Jeopardy." I should point out that that's IBM Watson, the program, not IBM Watson, the brand that came later. That's a whole other story. There really is a lot of hype there. Of course, there is success in robots. My point is there are a lot of impressive things going on. What's really interesting is that all these successes come from a very simple paradigm that's been remarkably successful.

This is probably my most technical slide. I want to show you how all this AI success is generated, even if you're not a computer scientist. It really comes from something called machine learning. I should caution you — our field, artificial intelligence, machine learning, is full of grandiose labels. Artificial intelligence is still not quite intelligence. Machine learning isn't exactly learning.

Let's talk about what it does do. Machine learning is, you take a bunch of categories pre-specified by a person. In my example, we have some cats. We've got some dogs. We've got some flying animals. Those are the categories. You take some examples of these categories. These are typically represented as vectors of numbers, and you associate a label with each vector. The vector says this is a cat. This is a dog.

When I say *you*, I should be clear. This is done manually. The computer is not doing that. The definition of the categories, the creation of the labeling of the data, is all done by people. There's a machine learning algorithm that does what I like to call the last mile of learning. After everything has been set up with copious manual labor by people, then a statistical algorithm looks at the labeled data and creates a statistical model that is then used to make predictions.

The next vector I see is a cat. The next image I see is a dog. This is the paradigm of machine learning. What's remarkable is that you can apply this paradigm to speech signal, you can apply it to email messages, categorize them as spam or not. You can apply it to road following, cars staying on the road — this is what's been driving all the AI successes that you've heard of. This is great stuff. If you ask me what's going to happen over the next five or ten years, we're going to see more and more of this in lots of arenas, in healthcare, in enterprise, in all kinds of places.

At the same time, we have to remember that this is still a very limited technology. These capabilities are very, very narrow. If we can do chess, we can't do other things. The program that plays Go can't do other things. We'll talk about that. I like to call these things AI savants, because they're so, so narrow in their capabilities. The poster child of this — on the one hand, tremendous but at the same time narrow capabilities — is, of course, AlphaGo, the program that in March 2016 defeated the world champion in the ancient game of Go, which is a fantastically difficult game with many, many options.

The thing is it's still a very, very limited system. I imagine to myself, let's say, I go to some AI cocktail party, the kind of place where geeks go. Let's say I meet AlphaGo and we were to have a little dialogue. I might say 'AlphaGo, congratulations! You defeated Lee Sedol, but can you play poker?' AlphaGo would say, 'No. I can't.' I'd say, 'Can you cross the street?' AlphaGo would say, 'No. I can't do that either.' I'd say, 'Can you at least tell me something about the game of Go,

the game that you won?' AlphaGo would say, 'No.' It can't explain itself. The remarkable thing is AlphaGo doesn't even know it won. All it is is a fancy calculator applied not to multiplication and division but to analyzing the different possibilities in the game of Go, which, of course, is a game that's black and white with moves. It's very, very artificial and stylized.

What we need to understand is programs like AlphaGo are intelligent in a sense, a narrow sense — it's a savant — but what they *aren't* is autonomous. This distinction is really, really important because we're used to thinking of intelligence and autonomy as going hand in hand. The intelligence that we're familiar with is intelligence of people, and people are, of course, autonomous as well. With machines, it's really, really different. To make this distinction clear, I'm setting up a two-by-two table so we can analyze this. What I'm trying to show you here is that intelligence is not autonomy. These are orthogonal notions.

"It's very scary to have a weapon that can make a life or death decision without a human in the loop. The intelligence is not the problem. It's the autonomy."

Let's look at the upper left-hand corner here. What we have is high autonomy with low intelligence — that's a bunch of teenagers drinking on a Saturday night. In contrast, on the bottom right, we can have high intelligence with very low, minimal, essentially no autonomy. That's a program like AlphaGo. Its autonomy is restricted to choosing which move to make in the game. It can't even decide on its own to play another game. That's how limited it is. This distinction is really important.

Let's look at various other examples and remember it. Let's take, for example, computer viruses. They're dangerous, potentially very destructive, but their danger comes from their autonomy, from the fact that they can go from one machine to the other over the internet, duplicate themselves, potentially causing massive damage. They're not intelligent.

I like to point out that my seven-year-old is really more autonomous than any AI system. He's very intelligent too, but that's not my point. He can cross the street. He can speak English. He can hear English, at least when he wants to. He doesn't play Go, but he's a highly autonomous being — again, the opposite of this savant-type behavior we see in AI systems.

Let's talk about AI weapons. That's a topic that has received some attention. Often it gets people's heart racing. You start thinking about 'Do I really want intelligent weapons?' Imagine a weapon that can launch itself, fly halfway across the world, and kill somebody. That's the stuff of nightmares.

What I want to point out is that the nightmare has to do with the autonomy. It's very scary to have a weapon that can make a life or death decision without a human in the loop. The intelligence is not the problem. It's the autonomy. Intelligence in weapons could actually prevent mistakes like we've had where innocent civilians get killed. So, the key thing that we want to avoid, again, is not intelligent weapons but autonomous ones, or at least we want to think about very carefully. Again, I'm just giving these examples to set up this key distinction.

To those of you who are very optimistic about AI, I do want to pose the question about what's AI's impact going to be on jobs? That's a very serious question that merits a lot of discussions that I don't necessarily have the answer for. Hal Varian, the chief economist at Google, has said that old jobs are going away but new jobs are going to come and replace them. I, frankly, don't think it's that simple. It's a question of, at what rate? Old jobs have gone away. New jobs have come. But it seems like this change is happening at an unprecedented rate.

I'm not a Luddite. I'm not suggesting that we just stop AI. These are pictures of the Luddites throwing shoes into the looms because they were concerned about their jobs in the textile industry. Obviously, the progression of technology has a lot of benefits. Think about washing machines, antibiotics, textiles — it's not something that we can just rule one way or another. Still, people ask the question, 'If it's not clear what impact AI will have on our society, if there are pluses and minuses, why don't we at least declare a moratorium on AI? Why don't we slow down and give ourselves a chance to think about it?' That can be quite appealing.

"... if we slow down AI progress in this country, we very much do so at our peril. Right now, we have something of an edge."

Bill Gates himself said — when you come from Seattle, you say things like 'Bill Gates himself.' This is almost like a deity. Bill Gates is a very smart guy. He said, 'Why don't we put a tax on robots?' Which, of course, has the effect of slowing things down, at least the proliferation of robots. The problem I have with that idea is that AI is very much a global phenomenon. China has declared explicitly that they want to be the world leader in AI by 2030. It's right around the corner. Putin has said that the leader in artificial intelligence will rule the world.

So, if we slow down AI progress in this country, we very much do so at our peril. Right now, we have something of an edge. I'm not sure that we want to give that up. Even in the case of weapons, which is a very, again, tricky issue, I don't mean to simplify it. But when people ask me about AI weapons, I say the one thing that I fear more than highly powerful AI weapons in the hands of our

military is highly powerful AI weapons in the hands of rogue nations, in the hands of terrorists. We actually benefit from a healthy competition in this area.

There's a whole other dimension to this, though, and that's the dimension that Paul Allen, my boss, had in mind when he created the Allen Institute for Artificial Intelligence. That's to set up a mission of AI for the common good. So, at Allen AI, or AI2, as we call ourselves, we're not trying to use AI to create weapons, to violate people's privacy, not even to sell you things. We're really trying to use AI to make the world a better place. I want to give you two key examples just to highlight the potential benefits of AI; one example we're working on, and the second example we aren't, but many other people are.

The first example has to do with AI-based scientific breakthroughs. The number of papers — and I think I don't need to tell anybody in this audience this fact — the number of scientific papers is growing explosively. It's actually doubling every few years or so. There are thousands and thousands of papers on cancer alone being published every week. Let me tell you, they're not getting easier to read, either. The number of papers that any of us can read in our lifetime is relatively small, and the number of papers is growing. So, we have a problem. There are no Renaissance men and women anymore. What we need is some kind of tool to help us be better scientists, to help us be better engineers.

We have a project at the Allen Institute of AI called Semantic Scholar. What it's doing is using natural language processing, the ability for computers to understand certain things in text — not understand Shakespeare and nuance connotations but try to extract the basic facts from turgid scientific prose. Map that into knowledge that the computer can use and that scientists can use. They don't have to read all these thousands and thousands of papers.

"'What if the cure for an intractable cancer is right now hidden in all these thousands, if not hundreds of thousands, of different papers and different studies? Can machines help us by teasing apart this information, synthesizing it, and presenting it to a medical researcher to help him or her make progress on that cancer?'"

Without going into the technical details here, our vision is to say, 'What if the cure for an intractable cancer is right now hidden in all these thousands, if not hundreds of thousands, of different papers and different studies? Can machines help us by teasing apart this information,

synthesizing it, and presenting it to a medical researcher to help him or her make progress on that cancer?' That's some of the potential that we're working on every day at the Allen Institute for AI.

My colleague, Eric Horvitz, likes to say it's the absence of AI technologies that's already killing people. It's not that AI is going to be terminal-like and kill us. It's quite the opposite. He's not just talking about Semantic Scholar and the information in scientific text. The third-leading cause of death in American hospitals is some kind of doctor error. Information systems, AI systems that can analyze what's happening in a hospital, that can detect potential mistakes that exhausted and overworked doctors make, could really save an enormous number of lives.

A whole other arena, this is one that we're not working on, is driving. Frankly, human drivers worry me. I have a 17-year-old. He's texting and driving. I know this because he's in the car and I get a text from him. I have this moral dilemma — do I text him back and say stop texting or do I not text because I don't want him to read my text? You worry about your kids. They're texting and driving. Even if your kids aren't texting and driving, other people's kids are and they're going to run into your kids. It's a huge problem. Certainly, some of his classmates are drinking and driving. DUI is a major problem for us.

It doesn't even have to be this nefarious. A while ago a friend of mine in Seattle was jogging and he was hit by a car. Thankfully, he's okay, the car was going very slowly. It was driven by a 96-year-old driver. It was not a great idea, but, at the same time, as we get older and we have parents, we don't want to limit their mobility. We want them to continue to be independent and so on, but how do you do that and still keep the rest of us safe? Obviously, self-driving technology can make a huge difference here, a very positive one. There are more than 30,000 highway deaths each year, close to a million accidents. A lot of these can be prevented if we have better technology using AI techniques.

Again, remember, AI is a tool here. It's really not that different than anti-lock brakes or automatic transmission. It's not like these cars that are going to become increasingly safe. It's not like a hundred cars are going to band together and say we're going to take over the White House. They don't decide where to go. They're tools at our disposal. It's important to remember that, again, this distinction between intelligence and capability, safe driving and autonomy, which still remains in our hands.

The question then becomes, 'Am I suggesting that everything's fine? AI's going to be beneficial. Sure we have to worry about jobs, but we'll leave that to the economies. But let AI blossom unconstrained, unfettered.' No, I'm not saying that either. It's a complex issue. We need to think

about how we prevent AI from harming us. What are negative uses of AI? When you think about that, there's actually been relatively little written or thought about this. One almost naturally goes to Asimov's Three Laws of Robotics. How many of you have read some of Asimov's stories? Okay, most of you but not everybody.

Let me quickly review, because even for you it's been a long time. His three laws are: 1) A robot may not allow a human being to come to harm — it seems like a good idea; 2) A robot must obey its orders so long as it doesn't conflict with the first law; so you can't order a robot to harm somebody; and then 3) A robot should protect itself, again, so long as it doesn't conflict with the first or second laws. These laws are really quite elegant, and they've survived for more than 60-70 years. At the same time, they're quite ambiguous.

"... to understand what's harmful or what's not really requires common sense. Remarkably, that's been one of the hardest things for us to give to the machine."

Asimov had all these stories that you read where he showed that there's contradictions. There are problems. It's not simple to enforce or even understand what these laws are. It has a lot to do with, what is this notion of harm? You can say "harm," but how do you communicate that to a computer? All the computer understands is a programming language.

Imagine that I have Alexa on my laptop, or Cortana, or whatever it is. I tell it, 'Reduce utilization on my hard drive.' It says, 'Yes, Oren, done.' I say, 'What did you do?' It says, 'I took your dataset that took you years to assemble. It's not backed up anywhere. There's about 15 gig. I deleted it. That proposal you wrote for funding. I deleted that too and all the copies. I've saved you a lot of space.' So, it was obeying my command but in doing that, it created — I see people shuddering. It creates a huge amount of harm.

The problem is that to understand what's harmful or what's not really requires common sense. Remarkably, that's been one of the hardest things for us to give to the machine. There really are no machines today with even a modicum of common sense. I like to represent that with what I call the AI car wash. This is a guy washing his car in the rain. It's not a great move. This is what AI systems are like today. While they can play Go very well or even do speech recognition, they have no common sense. That's a huge problem if you want them to be able to do things and, at the same time, avoid harm.

Basically, Asimov's laws are fantastic, but they're not very practical. In trying to be more pragmatic about it, I wrote an Op-Ed piece for *The New York Times*. It just came out in September. I tried to

suggest a regulatory framework. Again, not all the answers by any means but some ideas for thought. How can we think about constraining AI? First of all, to those of you who raised your hands being scared of AI, I do believe that we ought to put an impregnable "Off" switch on any AI system. This is a picture from *2001: A Space Odyssey* where HAL, the computer, says, 'I can't do that, Dave.' When the computer is killing Dave and says, 'I can't do that,' we need to be able to just turn the computer off. That's a fundamental principle.

"The line between computer technology and AI technology is actually very unclear. Instead, my suggestion is, let's regulate AI applications. Let's not try to regulate the research."

Another thing, too, to think about is the fact that it is very hard to regulate or constrain the research field itself. It's fast moving. It's amorphous. The line between computer technology and AI technology is actually very unclear. Instead, my suggestion is, let's regulate AI applications. Let's not try to regulate the research. When we have cars that have AI and, of course, have the potential to cause accidents, when we have toys that have AI and have the potential to violate privacy, when we have robots in the workplace and potentially in the home, those are things that we can regulate, and we should.

"We should have a rule that a computer system should engage in full disclosure and disclose that it's an AI system ... we don't want our AI that's privy to more and more information about us to reveal that information."

What might that look like? The first thing we have to adopt is the notion of responsibility. An AI system ought to be subject to all the laws that apply to its human operator, its human manufacturer. If my AI car crashes into yours, I can't say, 'Don't blame me. It was my car.' 'My AI did it' is not an excuse. We have to take responsibility for our intelligent cars the same way we have to take responsibility for our unintelligent cars. That's an important legal principle. It needs to be elaborated, but it will help prevent irresponsible use of AI.

A second thing that's becoming increasingly important is full disclosure. It's easier and easier. That's going to change even more so in the future for an AI to pretend that it's actually a person. We see this in Facebook. We see the Twitter bots. We now found out that in the most recent election there was more and more of bot activity masquerading as human activity, etc. We should have a rule that a computer system should engage in full disclosure and disclose that it's an AI system.

Another remarkable thing that's happening is our technology is violating our privacy more and more. That's even before we have AI. This is from a recent article. It turns out that, in certain instances, Google actually gives a user a number of options, including saying, 'I don't like this ad because it knows too much about me.' This is not science fiction, this is not a proposal, this is a real thing that Google has unveiled because some people are just getting more and more upset by how much our technology knows about them.

If you think of systems like the Amazon Echo, which records audio in your house. AI Barbie — we have these Barbie dolls with chips in them engaging in dialogues with our kids. Who knows what information our kids are telling Barbie? I find that more scary than funny. Even the Roomba — you think, 'I've got this little hockey puck robot. It's cleaning dust in my house. What's the big deal?' It turns out that in the process of doing that, it's building a map of your house. Apparently, iRobot, the company that manufactures this, was actually considering selling this information to third parties. They didn't do it, but they were considering it. It's not something that you thought that your AI robot was doing.

"We want our AI to avoid bias. AI actually has a great potential to catch human bias, whether it's in judges or in people to alert us to, say, loan processing gone awry."

Imagine more sophisticated robots. They pick up the phone and say, 'He needs a new carpet. It's really fraying on the edges. Help me out here.' This is not something you want happening without your approval. So, we don't want our AI that's privy to more and more information about us to reveal that information.

Another topic — this I have to confess wasn't covered in my Op-Ed, partially for space reasons — I'm adding one more regulatory principle. We want our AI to avoid bias. AI actually has a great potential to catch human bias, whether it's in judges or in people to alert us to, say, loan processing gone awry. We have studies that showed that when judges get hungry, they tend to produce more negative decisions. So, AI could alert to, 'Your Honor, maybe it's time for a snack, your blood sugar is dropping.' There's a lot of benefit here.

There's a really interesting problem. This is a bit technical with AI. It goes like this. The data that we give our machine learning system is typically culled from the real world. It may contain some bias in it, all kinds of bias. What's remarkable about machine learning technology is that it tends to generalize. It attempts to compress the data into some general principle. When it does that, it can actually amplify the bias that's in the training data, which is very negative. So, whatever bias we have in the training data, let's say loans were denied to people of a certain gender or a certain

race some percentage of the time, the last thing we want is AI to automatically say, 'I get the pattern here. Let's be more aggressive on denying that. That's what my data is telling me.'

This, by the way, is a problem that does have a technical answer. One of the research scientists at AI2 just won a best paper award on work to avoid amplifying bias even beyond what's in the training data. That's another important issue. We would agree, we want to avoid bias.

I could go on and on, but I want to leave us time for questions and discussion. I just want to conclude by highlighting what I consider the most important point. I started with this question: Is AI good or evil? My answer is it's neither. It's neither good nor evil. It's a tool. It's a technology. More than anything, it's a pencil. It's a fancy pencil, one that we can draw amazing pictures with. But a pencil is a tool that we use. We get to choose. Do we draw nice pictures, or do we draw unpleasant, horrific pictures? The choice is ours. I hope you'll join me both in the conversation and in working as a society to make sure that AI is used for good, not for evil.

Thank you very much.

"Is AI good or evil? My answer is it's neither. It's neither good nor evil. It's a tool. It's a technology. More than anything, it's a pencil. It's a fancy pencil, one that we can draw amazing pictures with. But a pencil is a tool that we use."