“Storytelling Teaches Robots Right and Wrong” Excerpt Transcript

Excerpt from February 26, 2016 episode of Science Friday.

IRA FLATOW	This is Science Friday. I'm Ira Flatow. Kids spend years and years by your side. They're in school learning the rules of the game, how our culture works, how to be polite, the differences between right and wrong. But robots? We're seeing more and more of them everywhere you look. And almost all of them are born with sociopathic tendencies, according to my next guest. They know only what their instructions tell them to do, which is usually accomplish a task as quickly and efficiently as possible. So how do you teach these robots how to behave ethically with good etiquette? Mark Riedl is an associate professor in the School of Interactive Computing at Georgia Tech in Atlanta. Welcome back to Science Friday.
MARK RIEDL	Hi. It's a pleasure to be here.
FLATOW	Well, what is this all about? The last time we talked, you were training robots to learn story plots and write fiction. And now you're teaching them to behave like good citizens. And you're doing this by analyzing stories, copying the protagonist behavior, is that right?
RIEDL	Yeah, that's right. So what we've been seeing over the last few years is really this kind of keen interest in the question of what happens when robots and other artificial intelligence agents, like Siri and Cortana, really start engaging with us on a more of a day-to-day, kind of social level. And the question arises, are these going to be safe? And I don't just mean in terms of physical damage or physical violence, but in terms of disrupting the social harmonies, in terms of cutting in line with us or insulting us. And what happens is when robots are trying to super optimize, they can do this quite unintentionally. So we wanted to try to teach agents the social conventions, the social norms, that we've all grown up with and we all used to get along with each other. And the question became, how do we get this information into the computer, into the robots? They don't have an entire lifetime to grow up and to share these experiences with us. And we looked at stories. And stories are a great way, a great medium, for communicating social values. People who write and tell stories really cannot help

	but to instill, in most parts, the values that we cherish, the values that we possess, the values we wish to see in good upstanding citizens in the protagonists, in the people that we talk about when we tell stories.
FLATOW	Can you give me an idea of the stories that you're reading to them or telling them?
RIEDL	Well there's kind of two different parts to this. There's kind of the long term vision where we imagine feeding entire sets of stories that might have been created by an entire culture, or an entire society, into a computer and having them reverse engineer the values out. So this could be everything from the stories we see on TV, in the movies and the book we read, really kind of the popular fiction that we see.
FLATOW	But you must have to pick and choose between the kinds of moral stories and ethical dilemmas that you want. Because if you sit there and watch television, you're going to think everything is about shooting and cups and robbers.
RIEDL	Well, what machine learning systems do, what artificial intelligence really is good at doing, is picking out the most prevalent signals, the most prevalent parts. So the things that it sees over and over and over again, time and time again, are the things that are going to kind of rise and bubble up to the top. And these tend to be the values that the culture and the society cherish. So even though we see, say, TV shows now with antiheroes, that's still a vast minority of the things that our entire society has output. So it's going to pick up on the most common values that it sees.
FLATOW	And so what do you train the robots or the computers in them to do? Give us an idea.
RIEDL	Right. So at this stage of things, I'm working on a software artificial intelligence system called Quixote. And we're still in very early stages of the research. So what we do now is we're still giving the system very simple stories, very childlike stories, really kind of procedures. And they're somewhat literal, you know. John walks to the bank. John stands in line waiting for the teller. Because those are stories that are a little bit easier to understand. But what it turns out is, when you ask people to tell stories even that simple, they tend to tell the computer what should happen, what the computer should expect to happen along the way. And, again, these kind of encode the quote unquote "right" way of doing things. Meaning that they don't steal, they stand in line, they're polite to people, so on and so forth.

FLATOW	So you're actually putting your moral and ethical values into the robot.
RIEDL	Well whomever we choose to tell the stories to the robot. So we chose stories not only because they're very good at conveying the social values. But also stories are very easy to tell. So virtually anyone could program a robot or program an agent in AI simply by telling stories. We don't have to teach people how to teach computers.
FLATOW	In the movie Ex Machina, Ava, the robot, is trained by using big data, big Google searches. Of course, spoiler alert, the outcome is not so good. Did you consider this kind of method instead of just storytelling?
RIEDL	Well, what we're actually doing is working towards a big data approach. But instead of taking all the data in the world, what we want to say is, let's focus on stories, because those are where the values of a culture are going to manifest themselves more correctly. Stories have everything from meeting with another character in a restaurant, and the social protocol you go through there, to exemplifying the moral standards of, let's say, trying to save the universe. When you think of protagonists as being those characters that exemplify the things that we see as the best parts of our society, antagonists, or the bad guys, are the opposite, and they kind of get their just desserts. So you should be able to, in theory, learn that certain behaviors lead to negative outcomes.
FLATOW	We wish you great success with it, and come back and tell us about your results.
RIEDL	I'd look forward to doing that.
FLATOW	Mark Riedl, associate professor in the School of Interactive Computing at Georgia Tech.