In Pursuit of the Ethical AI

Speaker:

Francesca Rossi, IBM T.J. Watson Research Center

Interviewer:

Jennifer Schenker, The Innovator

(Transcription by RA Fisher Ink)

Schenker: So, this morning we're here to talk about ethics and AI. I think that people in this room would all probably agree that if AI can help with things like pattern recognition; to eradicate disease, and help us solve problems that humans just haven't been able to solve until now, that we have no choice but to embrace it. However, when we start to get into autonomous systems, autonomous cars that literally make life and death decisions about who should live, who should die. That's when people start getting uncomfortable, and not just people who don't understand technology, but people who really do. Elon Musk, Bill Gates, Steven Hawking have all expressed real concerns about a lack of governance on AI, and have raised questions about the ethical challenges that AI poses. So as a result of that a number of industry groups have formed recently.

Elon Musk has launched the "OpenAI" Institute. A partnership on AI has been formed by Amazon, Deep Mine, Google, Facebook, IBM, and Microsoft. Stanford has launched a 100-year study on AI's impact on society. The White House has recently released a report on the future of AI which suggests that schools and universities should all be teaching affects in courses on AI, machine learning, computer science. In addition to all of these industry and government efforts, there's actually some initiatives that are underway to try and codify human values into AI systems.

Here to tell us about that is Francesca Rossi, and she is uniquely qualified to be discussing this topic. She wears many, many hats so I'm not going to try to mention all of her accomplishments and roles. But I think it's important to mention a couple of them. She is a research scientist at IBM's TJ Watson Research Center, professor of Computer Science at the University of Padova, Italy. A big part of her focus, is on ethical issues surrounding the development of behavior of AI systems. In particular, decision support systems for group decision making, and that's something we will come back to. She's co-chair of the Association for the Advancement of Artificial Intelligence AI and Ethics Committee, a member of the executive committee of the IEEE's global initiative on ethical considerations on the development of autonomous intelligence systems, and she's a member of the World Economic Forum's Council on AI and robotics. So tell us, Francesca, from a philosophical point of view, who decides what are human values? Is it a programmer rushing to meet some sort of product release deadline? Is it a company that has to report to shareholders? Should it be the

government? Should it be a standards body? How do we define, "What are human values?" How do we have agreement on that, and who decides?

Rossi: Yes, so that's very good question, there are not many answers to that. I think the only way to tackle that question is to really engage in a very multi-disciplinary discussion with people and experts from many disciplines, not just AI people like I am, but also philosophers, psychologists, and many different disciplines. In fact, it's one of the first questions that comes when you start discussing these issues, both in this IEEE initiative that you mentioned, but also when you want really to do solid and concrete research to actually embed these ethical principles into, for example, decision support systems, like in my case. For example, I had a project that has to do exactly with data, and where we work, AI people together with psychologists from Harvard, for example, who study actively in this discussion. Joshua Green, some of you may know him, and he's really great in trying to understand what values should be put in. Of course, there are many initiatives that think that somebody has kind of a cloud sourcing approach, like at MIT, some people use that to understand what people really think about, for example one instance of these moral dilemmas that these people have talked about, like The Trolley Problem and self-driving cars. But, you can have many other instances of these morally related decisions that AI systems will have to make.

Really the point is that I think there is definitely no universal set of values, it's very task specific and scenario specific. So, the kind of ethical principles that you want to embed in systems for a companion robot are certain, and for self-driving cars there's another one. For application in the healthcare domain, to help doctors find the best diagnosis, the best therapy, there's another one. It's very task specific, and I think that that's the only constructive way to tackle and try to give answers to this question. I think that we have to take a really constructive and even start from narrow AI, very task specific approaches. This is also the very constructive approach, this is also the approach that some of the people you mentioned, like Elon Musk, is taking. As you know, it's true that he expressed concerns, but also, he supported the Future of Life Institute with a very generous donation to fund research efforts, my project is one of them, on actually understanding, with researchers and people from other disciplines, how to tackle these issues. I think that's the best way.

Schenker: It is really hard to tackle this issue, because not only is there not agreement on human values as you mentioned. I'm not sure if you ask people, "What are your values?", that they even know. And even if they profess a certain set of values, humans are not consistent. They don't always carry out what they say is important to them.

Rossi: And in fact, in the discussion it came out that we think that machines, once we understand how to embed ethical principles into intelligent machines, could also help humans in this respect. Alert about ethical deviations that we are not going through the right path. Think about a hiring committee that is trying to choose among candidates, and this decision support system being there with this group of people and seeing that something in the decision making process is not done in the right way according to the professional code, or the social norms, or

the rules that you should follow. That could help people, individuals, or groups, to be more ethical than we usually are.

Schenker: So, there's been some concern expressed about the idea of Silicon Valley being the mission control for humanity, and so as you mentioned there are groups forming that bring in a lot of different players coming from different perspectives to help us figure all of this out. One of them, as you also mentioned, is at MIT. It's called "Society-in-the-Loop", and the idea is that you poll the public on decisions, and then you train AI to behave in ways that fairly reflect the values of people in much the same way that elected officials are supposed to represent our point of view. So, how well does this work in practice? Are there other ways to come to group decision making?

Rossi: So that's an interesting idea, that was in the nice experiments that they did. They are now collecting the data that came out of the many people who participated in this, responding to these questions about morals, how people would behave within a moral dilemma. But I think we're still at the beginning of this very interesting discussion, and I think that this fear that there is this governance or rules imposed by Silicon Valley. One thing that you mentioned is this partnership for AI, that we put together five of the main players in actually developing, and deploying AI into the real world, because we really believe that we need to understand these issues that only those that deploy AI to the real world can understand. And once you understand the issues, then we can discuss it, but not discuss it among the five companies that founded this initiative but among everybody, all the stakeholders from users, to developers, to scientists, to AI people, to non-profit organizations, to policy makers, to scientific associations. So, the partnership that has been formed is actually a place for everybody to have a voice in the discussion but especially those that deploy AI in the real world, because that's where the discussion should start.

Schenker: Some people envision a future in which every AI will have its own ethics module. It will be like a kind of ethics API that could be adapted to different professions and real life scenarios and dynamically adapted over time. How would that work from a technical perspective? Would it combine both rule-based and machine learning approaches? How could you actually create an ethics API?

Rossi: So there are these two main approaches that people have been discussing. The rule-based approach which is kind of a top-down approach, it's like the expert system, old approach, and also you understand more or less what are the rules that you have to follow to function according to certain moral values and ethical principles. You code them up into rules, and once they are rules you can code them into the machine. But of course it's very brittle, it doesn't cope well with the uncertainty of real-world scenarios and is not easy generalizable, so that cannot be enough. It could be for some aspects of these ethical principles, but it cannot be enough.

Then there is this other bottom-up approach that says, "Okay, let's just let the machine observe what humans do in their everyday profession," assuming that they perform their job in an

ethical way, "and then let's just observe, and by some machine learning technique, like reverse reinforcement learning or other ones. Let's just build the machine that actually can behave without really having rules on how to do that, and can behave in an ethical way." Of course, this is another completely different approach from the rule-based ones, but again, this is not enough by itself, we think, because by just observing and behaving like people, you may leave out some very essential behaviors. People usually mention this chat bot that Microsoft put over Twitter that just observed and behaved like the other people, and soon it became not very ethical. You have to really put the two approaches together. The whole scientific and research discussion is on how to put these two things together, how to combine them in order to make something which is general enough as the main rules, but is also flexible and nonrestrictive.

Schenker: I think you've made a really interesting observation. We always assume that we're the gold standard, but in fact people are not necessarily very ethical. So, if we're moving towards a world of autonomous systems, but also AI and people will be jointly making decisions. Say the AI is scanning millions of x-rays, or whatever it is, looking for anomalies and then the human medical expert looks at the difficult cases and makes a decision. Will autonomous and semi-autonomous AI require different kinds of restraints?

Rossi: I mean, of course there are different issues to be considered, but I think that this issue of making sure that intelligent machines behave according to some ethical principles that are aligned to our values is true for both kinds of systems. Of course if it's autonomous and you delegate decisions to these systems, you want to make sure that it knows what is a good or a bad decision. People understand that in an autonomous system that's a very crucial thing. But also in decision support systems, which is my main area of research, where I see the human in the loop, and actually in a symbiotic system between the machine and the human where the final decision maker is the human, but the system is either helping the human make a better decision. Even in those systems, if you don't have the right level of trust from the human to the AI system, then you cannot fully take advantage of what this AI system can give you. You don't want the human to over-trust the system, because you want to know whether the system has limitations, and it has limitation usually, like it could have data bias issues or many other algorithmic limitations. The human has to be aware of that, to really understand what limits there are there and not take everything that the system suggests, but you don't want the human to under-trust the system. You want to have the correct level of trust, and you want to build that, of course, over time; over repeated interactions, not just one-shot things. We think in that respect, it can be very helpful, these capabilities of the machine and the human to really interact naturally with each other, and also for the system to explain why, in some form of explanation, to explain why it is suggesting certain decisions and not others, certain therapies, and not other ones. This is a very important interaction over time, context awareness and memory capabilities, that the system should have to build this correct level of trust, and without this trust, we will not be able to put together in the best way human and machines.

One more thing about this human plus machine, it has been recognized in many scenarios that actually this system of human plus machine performs much better than the human alone or the

machine alone. You mentioned the White House also thought about these issues and the human-machine collaboration. In one of the documents that was released by the White House a month ago there is a very nice example in a breast cancer discovery scenario, where the best pathologists have an error rate of 3.5%, the best AI system has an error rate of 7.5%. So if you stop there, you say, "Ah, okay. The AI is not doing as good as the human, so let's forget about AI in this kind of scenario." But actually, if you put together the best pathologists with the best AI system, the error rate goes down to 0.5%. So it means that really humans and machines have very complimentary and intelligent kinds of skills. They can combine with each other very well, and that is what we have to focus on.

Schenker: Okay, great. Now I'm conscious of time and I want to make sure that we have time for questions from the audience. Sir.

Q1: So a lot of the difficult ethical issues are not good vs bad but rather conflicts between ethical goods. Should I help many people a little bit or a few people a lot? Is that the kind of thing you think that invariably will fall humans, still not machines? I'm wondering what you think of that.

Rossi: Well, there are many moral dilemmas. We have to understand when it makes sense to delegate these moral dilemmas to machines, and when it does not make sense. But the main point is that we should try to understand how to code moral values, general moral values and ethical principles into these machines. Then, given that the machines have the ability to distinguish between good and bad decisions, but of course, with the uncertainty that we sometimes have, then we have to make decisions about what is reasonable to delegate to them and what is not. In some cases, we don't have a choice, because there is not enough time for the human to be in the loop. But in most cases actually there is time to be in the loop, like in health care scenarios. So in that case, the human should be the final decision makers, but it should be convinced, not just because of some manipulation or deception, but because of repeated interactions over time with the machine, that the machine functions with the same moral values as the doctor, for example. It follows the same Hippocratic Oath and all the other social norms and rules that the doctor would usually follow.

Schenker: We have time for one more question.

Q2: So one of the biggest areas of autonomy development is in the military sector, and I wonder if you're aware of any nascent attempts to establish some ethical standards in the military domain, because it's almost inevitable that we're going to have issues there.

Rossi: Well, there are many people that have very different opinions about that kind of development, of course. There are people that propose a ban on autonomous weapons, meaning that we should not delegate that kind of decision to machines. There are people that propose a moratorium on the use of those. I think there is a very active discussion that should again involve everybody to make us understand what's the best thing to do, for all of us, for humanity, for people. It's not true that we should just do whatever is possible to do. We should

think about it. So that's the main point in this discussion about AI ethics. We should care about what the limits are of what we can do with our technology, but we should also think about what is the world where our technology is leading, and discuss that, and decide exactly what is the future that we want for us and our children.

Schenker: On that note, we're going to have to wrap up the session. Thank you, Francesca.