AI Is the Future, So How Do We Create a Trustworthy AI?

I was the ghostwriter of this 2022 blog post for a high-ranking official in the tech space

Artificial intelligence used to be mainly a tool for high-level STEM research, but soon the technology made its way into the consumer realm. Now, AI is driving everything from resume screening to the content you see in web search results or on your favorite social media platform.

In the U.S., without any oversight from the federal government, private companies are using AI software to make decisions about employment, creditworthiness, health and medicine, and criminal justice—and they don't have to report to anyone about how they ensure that AI programs aren't encoded, either consciously or unconsciously, with structural biases.

"Part of the appeal of algorithmic decision-making is that it seems to offer at objective way of overcoming human subjectivity, bias, and prejudice," said Michael Sandel, Anne T. and Robert M. Bass Professor of Government at Harvard University. "But we are discovering that many of the algorithms that decide who should get parole, for example, or who should be presented with employment opportunities or housing … replicate and embed the biases that already exist in our society."

AI has tremendous potential benefits. It makes it possible to look at patterns much faster than humans can, and when properly calibrated, AI-driven applicant tracking systems can actually allow larger pools of job applicants without gatekeeping or favoritism, for example. But do we as technologists have the sophistication of addressing the moral and ethical issues of what's good and bad?

Let's say we as humans say, "we want people to be happy, and with artificial intelligence, we should build systems for people to be happy." What does that mean? Happiness means different things for different people, but would an AI catch the subtleties of that or just say "most people say happiness means this," enact that "happiness solution," and find out that in fact it does not make everyone, or even most people, happy?

Essentially, what are the unintended consequences of artificial intelligence?

Europe has been at the forefront of considering the benefits and consequences of artificial intelligence. In 2019, the European Commission's High-Level Expert Group on AI put forth a series of ethics guidelines required for a "trustworthy AI." They are as follows:

1. **Human agency and oversight,** including fundamental rights, human agency, and human oversight.
2. **Technical robustness and safety**, including resilience to attack and security, fallback plan, and general safety, accuracy, reliability, and reproducibility.
3. **Privacy and data governance**, including respect for privacy, quality and integrity of data, and access to data.
4. **Transparency**, including traceability, explainability, and communication.
5. **Diversity, non-discrimination, and fairness**, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.

6. **Environmental and societal well-being**, including sustainability and environmental friendliness, social impact, society, and democracy.
7. **Accountability**, including auditability, minimization and reporting of negative impact, trade-offs, and redress.

The world is changing, and AI will become an ever-increasing presence in almost every part of our lives. AIs are busy crunching data for life-saving drugs and treatments and assisting researchers with formerly laborious and potentially error-ridden tasks. AIs are taking the rote jobs out of workers' days, leaving them free to do more important things.

If we someday do manage to create an artificial intelligence system that can do 80 percent of the labor people do, that's going to mean big changes. For the hundreds of millions of people who work in factories to support their families, for example, it would be a disaster. What are we going to do for all the people whose jobs were lost to an AI? How are we going to manage the communities that will be destroyed by the loss of employment? These are the big, big questions that need to be talked about.

We have a paradox in technology and IT. We think a lot about IT being hardware and software and tools, and we think a lot about the visible risks of technology. But we do not think as much about the invisible risks of innovation.

Innovation has an impact. The invention of the internal combustion engine, for example, had a global impact. One simple innovation resulted in suburbia, popular mass transportation, and a significant environmental and economic shift. And the innovation of AI has numerous implications as well. We need a lot of discussion at the national and international level about the moral and ethical issues people are going to confront with AI technology—both positive and negative.