Principles and design guidance for responsible AI

Eric CHANG, Microsoft Research Asia, China

WFEO

Artificial Intelligence or AI has become a widely used, catch-all phrase that pops up in technical discussions, science fiction, analyst reports, news and beyond. It seems to get applied to almost any technology that is ground breaking or futuristic. But most experts agree that AI can be precisely defined in terms that even non-technologists can understand.

The past few years have brought radical improvements in terms of capability and performance in big data, analytics and artificial intelligence. These improvements are being adopted at an unprecedented pace by industry, enabling company to develop machine learning (ML) models, bots, and artificial intelligence (AI) applications, for a wide variety of needs. Misguided implementations of these technology could lead to increased risk of litigation, create limiting new regulations, or have negative impact on company, the industry and society if the ethical development and implementation of these technologies are not considered.

Designing AI to be trustworthy requires creating solutions that reflect ethical principles that are deeply rooted in important and timeless values. As we've thought about it, we've focused on six principles that we believe should guide the development of AI. Specifically, AI systems should be fair, reliable and safe, private and secure, inclusive, transparent, and accountable. These principles are critical to addressing the societal impacts of AI and building trust as the technology becomes more and more a part of the products and services that people use at work and at home every day.

Every powerful new technology in history has brought with it certain risks. What do we need to do to ensure that AI only helps humans and never harms us?