EDITORIAL

AI, people, and society

Eric Horvitz is a technical fellow and director of Microsoft Research Labs. He is a past president of the Association for the Advancement of Artificial Intelligence (AAAI) and cofounded the One Hundred Year Study on AI at Stanford. horvitz@microsoft.com

In an essay about his science fiction, Isaac Asimov reflected that "it became very common…to picture robots as dangerous devices that invariably destroyed their creators." He rejected this view and formulated the "laws of robotics," aimed at ensuring the safety and benevolence of robotic systems. Asimov's stories about the relationship between people and robots were only a few years old when the phrase "artificial intelligence" (AI) was used for the first time in a 1955 proposal for a study on using computers to "…solve kinds of problems now reserved for humans." Over the half-century since that study, AI has matured into subdisciplines that have yielded a constellation of methods that enable perception, learning, reasoning, and natural language understanding.

Growing exuberance about AI has come in the wake of surprising jumps in the accuracy of machine pattern recognition using methods referred to as "deep learning." The advances have put new capabilities in the hands of consumers, including speech-to-speech translation and semi-autonomous driving. Yet, many hard challenges persist—and AI scientists remain mystified by numerous capabilities of human intellect.

"Excitement about AI has been tempered by concerns about potential downsides."

Excitement about AI has been tempered by concerns about potential downsides. Some fear the rise of superintelligences and the loss of control of AI systems, echoing themes from age-old stories. Others have focused on nearer-term issues, highlighting potential adverse outcomes. For example, data-fueled classifiers used to guide high-stakes decisions in health care and criminal justice may be influenced by biases buried deep in data sets, leading to unfair and inaccurate inferences. Other imminent concerns include legal and ethical issues regarding decisions made by autonomous systems, difficulties with explaining inferences, threats to civil liberties through new forms of surveillance, precision manipulation aimed at persuasion, criminal uses of AI, destabilizing influences in military applications, and the potential to displace workers from jobs and to amplify inequities in wealth.

As we push AI science forward, it will be critical to address the influences of AI on people and society, on short- and long-term scales. Valuable assessments and guidance can be developed through focused studies, monitoring, and analysis. The broad reach of AI's influences requires engagement with interdisciplinary groups, including computer scientists, social scientists, psychologists, economists, and lawyers. On longer-term issues, conversations are needed to bridge differences of opinion about the possibilities of superintelligence and malevolent AI. Promising directions include working to specify trajectories and outcomes, and engaging computer scientists and engineers with expertise in software verification, security, and principles of failsafe design.

The good news is that studies, programs, and projects have been organized. In 2008, a multimonth study on long-term AI futures was hosted by the Association for the Advancement of Artifical Intelligence, culminating in a meeting in Asilomar, California. That meeting inspired the One Hundred Year Study on AI at Stanford University, a project charged with organizing similar studies every 5 years for a century and beyond (the first report was released last year). Other recent efforts include workshops and studies hosted by the U.S. National Academies. Last April, a report was published on influences of automation on the U.S. workforce following a 2-year study. Earlier this year, representatives from industry, academia, and civil society formed a nonprofit organization called the Partnership on AI, aimed at recommending best practices for developing and fielding AI technologies.

Asimov concludes in his essay, "I could not bring myself to believe that if knowledge presented danger, the solution was ignorance. To me, it always seemed that the solution had to be wisdom. You did not refuse to look at danger, rather you learned how to handle it safely." Indeed, the path forward for AI should be guided by intellectual curiosity, care, and collaboration.

–**Eric Horvitz**

10.1126/science.aao2466

PHOTOS: (INSET) KIYOSHI TAKAHASE SEGUNDO/ALAMY STOCK PHOTO; (TOP RIGHT) DAN DELONG FOR MICROSOFT

Science

AI, people, and society

Eric Horvitz

Science **357** (6346), 7.
DOI: 10.1126/science.aao2466

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.