MAKING A.I. SAFE: Realizing the Promise, Controlling the Peril

Ray Kurzweil is the author of five books on artificial intelligence, including the recent *New York Times* best seller *How to Create a Mind*.

December 19, 2014
TIME

Stephen Hawking, the pre-eminent physicist, recently warned that artificial intelligence (AI), once it surpasses human intelligence, could pose a threat to the existence of human civilization. Elon Musk, the pioneer of digital money, private spaceflight and electric cars, has voiced similar concerns.

If AI becomes an existential threat, it won't be the first one. Humanity was introduced to existential risk when I was a child sitting under my desk during the civil-defense drills of the 1950s. Since then we have encountered comparable specters, like the possibility of a bioterrorist creating a new virus for which humankind has no defense. Technology has always been a double-edged sword, since fire kept us warm but also burned down our villages.

The typical dystopian futurist movie has one or two individuals or groups fighting for control of "the AI." Or we see the AI battling the humans for world domination. But this is not how AI is being integrated into the world today. AI is not in one or two hands; it's in 1 billion or 2 billion hands. A kid in Africa with a smartphone has more intelligent access to knowledge than the President of the United States had 20 years ago. As AI continues to get smarter, its use will only grow. Virtually everyone's mental capabilities will be enhanced by it within a decade.

We will still have conflicts among groups of people, each enhanced by AI. That is already the case. But we can take some comfort from a profound, exponential decrease in violence, as documented in Steven Pinker's 2011 book, The Better Angels of Our Nature: Why Violence Has Declined. According to Pinker, although the statistics vary somewhat from location to location, the rate of death in war is down hundredsfold compared with six centuries ago. Since that time, murders have declined tensfold. People are surprised by this. The impression that violence is on the rise results from another trend: exponentially better information about what is wrong with the world—another development aided by AI.

There are strategies we can deploy to keep emerging technologies like AI safe. Consider biotechnology, which is perhaps a couple of decades ahead of AI. A meeting called the Asilomar Conference on Recombinant DNA was organized in 1975 to assess its potential dangers and devise a strategy to keep the field safe. The resulting guidelines, which have been revised by the industry since then, have worked very well: there have been no significant problems, accidental or intentional, for the past 39 years. We are now seeing major advances in medical treatments reaching clinical practice and thus far none of the anticipated problems.

Consideration of ethical guidelines for AI goes back to Isaac Asimov's three laws of robotics, which appeared in his short story "Runaround" in 1942, eight years before Alan Turing introduced the field of AI in his 1950 paper "Computing Machinery and Intelligence." The median view of AI practitioners today is that we are still several decades from achieving human-level AI. I am more optimistic and put the date at 2029, but either way, we do have time to devise ethical standards.

There are efforts at universities and companies to develop AI safety strategies and guidelines, some of which are already in place. Similar to the Asilomar guidelines, one idea is to clearly define the mission of each AI program and to build in encrypted safeguards to prevent unauthorized uses.

Ultimately, the most important approach we can take to keep AI safe is to work on our human governance and social institutions. We are already a human-machine civilization. The best way to avoid destructive conflict in the future is to continue the advance of our social ideals, which has already greatly reduced violence.

AI today is advancing the diagnosis of disease, finding cures, developing renewable clean energy, helping to clean up the environment, providing high-quality education to people all over the world, helping the disabled (including providing Hawking's voice) and contributing in a myriad of other ways. We have the opportunity in the decades ahead to make major strides in addressing the grand challenges of humanity. AI will be the pivotal technology in achieving this progress. We have a moral imperative to realize this promise while controlling the peril. It won't be the first time we've succeeded in doing this.