Preventing an AI Apocalypse

By Seth Baum

NEW YORK – Recent advances in artificial intelligence have been nothing short of dramatic. AI is transforming nearly every sector of society, from transportation to medicine to defense. So it is worth considering what will happen when it becomes even more advanced than it already is.

The apocalyptic view is that AI-driven machines will outsmart humanity, take over the world, and kill us all. This scenario crops up often in science fiction, and it is easy enough to dismiss, given that humans remain firmly in control. But many AI experts take the apocalyptic perspective seriously, and they are right to do so. The rest of society should as well.

To understand what is at stake, consider the distinction between "narrow AI" and "artificial general intelligence" (AGI). Narrow AI can operate only in one or a few domains at a time, so while it may outperform humans in select tasks, it remains under human control.

AGI, by contrast, can reason across a wide range of domains, and thus could replicate many human intellectual skills, while retaining all of the advantages of computers, such as perfect memory recall. Run on sophisticated computer hardware, AGI could outpace human cognition. In fact, it is hard to conceive an upper limit for how advanced AGI could become.

As it stands, most AI is narrow. Indeed, even the most advanced current systems have only limited amounts of generality. For example, while Google DeepMind's AlphaZero system was able to master Go, chess, and shogi – making it more general than most other AI systems, which can be applied only to a single specific activity – it has still demonstrated capability only within the limited confines of certain highly structured board games.

Many knowledgeable people dismiss the prospect of advanced AGI. Some, such as Selmer Bringsjord of Renselaer Polytechnic Institute and Drew McDermott of Yale University, argue that it is impossible for AI to outsmart humanity. Others, such as Margaret A. Boden of the University of Sussex and Oren Etzioni of the Allen Institute for Artificial Intelligence, argue that human-level AI may be possible in the distant future, but that it is far too early to start worrying about it now.

These skeptics are not marginal figures, like the cranks who try to cast doubt on climate-change science. They are distinguished scholars in computer science and related fields, and their opinions must be taken seriously.

Yet other distinguished scholars – including David J. Chalmers of New York University, Yale University's Allan Dafoe and Stuart Russell of the University of California, Berkeley, Nick Bostrom of Oxford University, and Roman Yampolskiy of the University of Louisville – do worry that AGI could pose a serious or even existential threat to humanity. With experts lining up on both sides of the debate, the rest of us should keep an open mind.

Moreover, AGI is the focus of significant research and development. I recently completed a survey of AGI R&D projects, identifying 45 in 30 countries on six continents. Many active initiatives are based in major corporations such as Baidu, Facebook, Google, Microsoft, and Tencent, and in top universities

such as Carnegie Mellon, Harvard, and Stanford, as well as the Chinese Academy of Sciences. It would be unwise simply to assume that none of these projects will succeed.

Another way of thinking about the potential threat of AGI is to compare it to other catastrophic risks. In the 1990s, the US Congress saw fit to have NASA track large asteroids that could collide with the earth, even though the odds of that happening are around one in 5,000 per century. With AGI, the odds of a catastrophe over the upcoming century could be as high as one in a hundred, or even one in ten, judging by the pace of R&D and the strength of expert concern.

The question, then, is what to do about it. For starters, we need to ensure that R&D is conducted responsibly, safely, and ethically. This will require a deeper dialogue between those working in the AI field and policymakers, social scientists, and concerned citizens. Those in the field know the technology and will be the ones to design it according to agreed standards; but they must not decide alone what those standards will be. Many of the people developing AI applications are not accustomed to thinking about the social implications of their work. For that to change, they must be exposed to outside perspectives.

Policymakers also will have to grapple with AGI's international dimensions. Currently, the bulk of AGI R&D is carried out in the United States, Europe, and China, but much of the code is open source, meaning that the work potentially can be done from anywhere. So, establishing standards and ethical ground rules is ultimately a job for the entire international community, though the R&D hubs should take the lead.

Looking ahead, some efforts to address the risks posed by AGI can piggyback on policy initiatives already put in place for narrow AI, such as the new bipartisan AI Caucus launched by John Delaney, a Democratic congressman from Maryland. There are many opportunities for synergy between those working on near-term AI risks and those thinking about the long term.

But regardless of whether narrow AI and AGI are considered together or separately, what matters most is that we take constructive action now to minimize the risk of a catastrophe down the road. This is not a task that we can hope to complete at the last minute.

Seth Baum is the executive director of the Global Catastrophic Risk Institute (GCRI), a think tank focused on extreme global risks. He is also affiliated with the Blue Marble Space Institute of Science and the University of Cambridge Centre for the Study of Existential Risk.