Artificial Intelligence: Are We Prepared for the Tripwire?

Yijun Jiang (10/20/2015, STS.034, third draft of op-ed)

This summer, the fifth episode of the movie *the Terminator* was released and earned nearly $450 million worldwide at box offices. In the movie, scientists created Skynet, an artificial intelligence (AI), or simply speaking a smart computer system, which is so advanced that it has gained self-awareness and treats humans as its threat. Led by the savior John Conner, humans must fight against their own machines to prevent themselves from being extinct.

The movie, though fictional, calls for people's awareness on the possible prospectives of AI. When working for us rather than against us, an AI as powerful as Skynet has huge potentials dealing with some of the biggest challenges of mankind such as global warming, energy crisis and cancer treatment. But on the other hand, we face the risks of creating a real Skynet. Is such a super AI far away from us? Do we need to worry about Skynet at the moment?

The risks are real. The risks may be near.

In a survey conducted by the Swedish philosopher Nick Bostrom in 2013, hundreds of AI experts were asked to predict the advent of an AI whose intelligence was comparable with humans. According to statistics, by the year 2050, such an AI was believed to exist with a probability of 50%. Moreover, more than 75% researchers believed that once a human-like AI was created, within 30 years AIs would become far more advanced than humans in all aspects.

I have to admit that this schedule is so tight that it sounds ridiculous. Unsurprisingly, a lot of people both inside and outside the field, as well as many software companies, refuse to believe it. Paul Allan, the founder of Microsoft, claimed that this schedule might have underestimated the difficulties in current and future AI research.

Despite of the rapid research progress over the past few years, all the AIs that we are working with, such as Siri and Google Driverless Car, are still in an elementary stage. They are usually programmed to be specialized in only one job. Although they are much more efficient than us in this particular job, they lack the ability to “think”, that is, to learn from experiences, make judgments and self-improve beyond what software engineers have taught them.

Skynet, on the other hand, has these features. It actively learns from the outside world and updates itself. This behavior is exactly like the evolution of ourselves. The only difference is, the evolution of a powerful AI takes much less time than us due to its hardware advantage that we organisms do not possess. In an epoch when supercomputers can simulate the evolution of a galaxy over billions of years within days, tracing the history of life and biological intelligence in a short time is no longer unthinkable, once the AI knows how to start.

We have not yet reached this starting point. And researchers come across great difficulties trying to get there. But this does not mean that super AIs are far away. The fact is, the pace of human development is accelerating over the history, and this trend will keep on into the future. People in the 1500s would be surprised by the development in classical physics and maritime technology if they had time traveled to 200 years later. However, the same kind of surprise can be experienced merely over the recent 20 years, when inventions like the Internet, smart phones and new materials came to us one after another. The important lesson is, it is misleading to predict the future based on our current pace. Breakthroughs come much faster than we have often expected.

This accelerating trend applies equally well to the study of AI. As we gain more and more experience on elementary AIs, we are accelerating the advent of the breakthrough, where AIs acquire the general ability to evolve and soon become far superior to us in intelligence and all other aspects. As the writer Aaron Saenz described, it is just like mixing amino acids on early earth and suddenly creating life among them. In the words of specialists, we unexpectedly step onto the tripwire.

Then the possibility of such an AI turning against us becomes an issue. We are not at all experienced in how to make a super AI safe. And if we cannot figure that out before triggering the tripwire, we will be in big trouble.

Unfortunately, what we have done is far from enough. Super AI security is not an easy job and demands much serious effort, possibly even more than the development of AI itself. To prevent a Skynet we have to teach an AI to respect the values of human society, a task much easier said than done. Undoubtedly, technological as well as ethical issues will emerge and entangle in the research. Because of such difficulties, our pace in this field is much slower than the creation of new AIs.

The second reason is that, driven by profits, most AI companies are completely focused on developing better AIs, without paying any attention to this long-term risk. Research on super AI security does not get enough funding that matches its significance. Continuing such aggressive effort in current AI development may lead us to the tripwire unprepared, when it is already too late.

Since a super AI will eventually come to us, and possibly in the near future, the wise thing to do is to start thinking of its risks right now. The race between AI innovation and AI security is perhaps the most significant competition in human history, whose result will very likely lead us to either our paradise or our grave.

As a big fan of *the Terminator*, I am looking forward to its sixth episode. However, in real life there is no John Conner for us to put faith in. We need to be responsible for every step we make pushing forward artificial intelligence. And it is never too early to ask ourselves if we are prepared for the tripwire.