Artificial intelligence systems more apt to fail than to destroy

March 24 2015, by David Stauth

The most realistic risks about the dangers of artificial intelligence are basic mistakes, breakdowns and cyber attacks, an expert in the field says – more so than machines that become super powerful, run amok and try to destroy the human race.

Thomas Dietterich, president of the Association for the Advancement of Artificial Intelligence and a distinguished professor of computer science at Oregon State University, said that the recent contribution of $10 million by Elon Musk to the Future of Life Institute will help support some important and needed efforts to ensure AI safety.

But the real risks may not be as dramatic as some people visualize, he said.

"For a long time the risks of artificial intelligence have mostly been discussed in a few small, academic circles, and now they are getting some long-overdue attention," Dietterich said. "That attention, and funding to support it, is a very important step."

Dietterich's perspective of problems with AI, however, is a little more pedestrian than most – not so much that it will overwhelm humanity, but that like most complex engineered systems, it may not always work.

"We're now talking about doing some pretty difficult and exciting things with AI, such as automobiles that drive themselves, or robots that can effect rescues or operate weapons," Dietterich said. "These are high-

stakes tasks that will depend on enormously complex algorithms.

"The biggest risk is that those algorithms may not always work," he added. "We need to be conscious of this risk and create systems that can still function safely even when the AI components commit errors."

Dietterich said he considers machines becoming self-aware and trying to exterminate humans to be more science fiction than scientific fact. But to the extent that computer systems are given increasingly dangerous tasks, and asked to learn from and interpret their experiences, he says they may simply make mistakes.

"Computer systems can already beat humans at chess, but that doesn't mean they can't make a wrong move," he said. "They can reason, but that doesn't mean they always get the right answer. And they may be powerful, but that's not the same thing as saying they will develop superpowers."

More immediate and real risks, he said, will be to identify how mistakes might occur, and how to create systems that can help deal with, minimize or accommodate them.

Some of the most imminent threats computers will pose in a malicious sense, Dietterich said, will probably emerge as a result of cyber attacks. Humans with malicious intent using artificial intelligence and powerful computers to attack other computer systems are a real threat, he pointed out, and thus it would be a good place to focus some of the first work in this field.

That work should receive a significant boost from the recent grant, Dietterich said, which will facilitate research around the world via an open grants competition.

Provided by Oregon State University