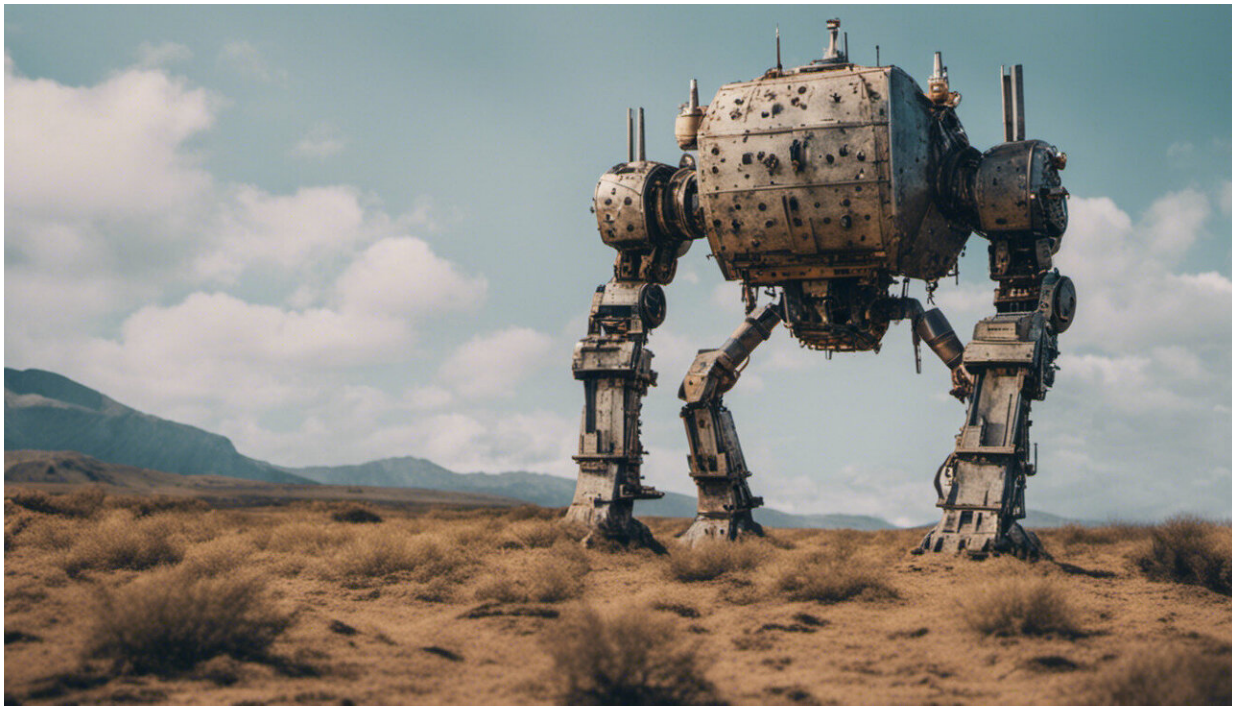
PHYS ORG

Never mind killer robots – even the good ones are scarily unpredictable

August 25 2017, by Taha Yasseri



Credit: AI-generated image (disclaimer)

The heads of more than 100 of the world's top artificial intelligence companies are very alarmed about the development of "killer robots". In an open letter to the UN, these business leaders – including Tesla's Elon Musk and the founders of Google's DeepMind AI firm – warned that autonomous weapon technology could be misused by terrorists and despots or hacked to perform in undesirable ways.

But the real threat is much bigger – and not just from human misconduct but from the machines themselves. The research into complex systems shows how behaviour can emerge that is much more unpredictable than the sum of individual actions. On one level this means human societies can behave very differently to what you might expect just looking at individual behaviour. But it can also apply to technology. Even ecosystems of relatively simple AI programs – what we call stupid, good bots – can surprise us, and even when the individual bots are behaving well.

The individual elements that make up complex systems, such as economic markets or global weather, tend not to interact in a simple linear way. This make these systems very hard to model and understand. For example, even after many years of climatology, it's still impossible to make long-term weather predictions. These systems are often very sensitive to small changes and can experience explosive feedback loops. It is also very difficult to know the precise state of such a system at any one time. All these things make these systems intrinsically unpredictable.

All these principles apply to large groups of individuals acting in their own way, whether that's human societies or groups of AI bots. My colleagues and I recently studied one type of a complex system that featured good bots used to automatically edit Wikipedia articles. These different bots are designed and exploited by Wikipedia's trusted human editors and their underlying software is open-source and available for anyone to study. Individually, they all have a common goal of improving the encyclopaedia. Yet their collective behaviour turns out to be surprisingly inefficient.

These Wikipedia bots work based on well-established rules and conventions, but because the website doesn't have a central management system there is no effective coordination between the people running different bots. As a result, we found pairs of bots that have been undoing each other's edits for several years without anyone noticing. And of course, because these bots lack any cognition, they didn't notice it either.

The bots are designed to speed up the editing process. But slight differences in the design of the bots or between people who use them can lead to a massive waste of resources in an ongoing "edit war" that would have been resolved much quicker with human editors.

We also found that the bots behaved differently in different language editions of Wikipedia. The rules are more or less the same, the goals are identical, the technology is similar. But in German Wikipedia, the collaboration between bots is much more efficient and productive compared to, for example, Portuguese Wikipedia. This can only be explained by the differences between the human editors who run these bots in different environments.

Exponential confusion

Wikipedia bots have very little autonomy and the system already operates very differently to the goals of individual bots. But the Wikimedia Foundation is planning to use AI that will give more autonomy to the bots. That will likely lead to even more unexpected behaviour.

Another example is what can happen when two bots designed to speak to humans interact with each other. We're no longer surprised by the answers given by artificial personal assistants such as the iPhone's Siri. But put several of these kind of chatbots together and they can quickly start acting in surprising ways, arguing and even insulting each other.

The bigger the system becomes and the more autonomous each bot is, the more complex and hence unpredictable the future behaviour of the system will be. Wikipedia is an example of large number of relatively simple bots. The chatbots example is a small number of rather sophisticated and creative bots – in both cases unexpected conflicts emerged. The complexity and therefore unpredictability increases exponentially as you add more and more individuals to the system. So in a future system with a large number of very sophisticated robots, the unexpected behaviour could go beyond our imagination.

Self-driving madness

For example, self-driving cars promise exciting advances in the efficiency and safety of road travel. But we don't yet know what will happen once we have a large, wild system of fully autonomous vehicles. They may well behave very differently to a small set of individual cars in a controlled environment. And even more unexpected behaviour might occur when driverless cars "trained" by different humans in different environments start interacting with each another.

Humans can adapt to new rules and conventions relatively quickly but can still have trouble switching between systems. This can be way more difficult for artificial agents. If a "German-trained" car was driving in Italy, for example, we just don't know how it would deal with the written rules and unwritten cultural conventions being followed by the many other "Italian-trained" cars. Something as common as crossing an intersection could become lethally risky because we just wouldn't know if the cars would interact as they were supposed to or whether they would do something completely unpredictable.

Now think of the killer robots that Elon Musk and his colleagues are worried about. A single killer robot could be very dangerous in wrong hands. But what about an unpredictable system of killer robots? I don't even want to think about it.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Never mind killer robots – even the good ones are scarily unpredictable (2017, August 25) retrieved 23 April 2024 from https://phys.org/news/2017-08-mind-killer-robots-good-scarily.html