PHYS.ORG

Fighting malevolent AI—artificial intelligence, meet cybersecurity

June 13 2016, by Roman V. Yampolskiy, University Of Louisville

Credit: AI-generated image (disclaimer)

With the appearance of robotic financial advisors, self-driving cars and personal digital assistants come many unresolved problems. We have already experienced market crashes caused by intelligent trading software, accidents caused by self-driving cars and hate speech from chat-bots that turned racist.

Today's narrowly focused artificial intelligence (AI) systems are good only at specific assigned tasks. Their failures are just a warning: Once humans develop general AI capable of accomplishing a much wider range of tasks, expressions of prejudice will be the least of our concerns. It is not easy to make a machine that can perceive, learn and synthesize information to accomplish a set of tasks. But making that machine safe as well as capable is much harder.

Our legal system lags hopelessly behind our technological abilities. The field of machine ethics is in its infancy. Even the basic problem of controlling intelligent machines is just now being recognized as a serious concern; many researchers are still skeptical that they could pose any danger at all.

Worse yet, the threat is vastly underappreciated. Of the roughly 10,000 researchers working on AI around the globe, only about 100 people – one percent – are fully immersed in studying how to address failures of multi-skilled AI systems. And only about a dozen of them have formal training in the relevant scientific fields – computer science, cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security and psychology. Very few are taking the approach I am: researching malevolent AI, systems that could harm humans and in the worst case completely obliterate our species.

AI safety

Studying AIs that go wrong is a lot like being a medical researcher discovering how diseases arise, how they are transmitted, and how they affect people. Of course the goal is not to spread disease, but rather to fight it.

From my background in computer security, I am applying techniques first developed by cybersecurity experts for use on software systems to this new domain of securing intelligent machines.

Last year I published a book, "Artificial Superintelligence: a Futuristic Approach," which is written as a general introduction to some of the most important subproblems in the new field of AI safety. It shows how ideas from cybersecurity can be applied in this new domain. For example, I describe how to contain a potentially dangerous AI: by treating it similarly to how we control invasive self-replicating computer viruses.

My own research into ways dangerous AI systems might emerge suggests that the science fiction trope of AIs and robots becoming self-aware and rebelling against humanity is perhaps the least likely type of this problem. Much more likely causes are deliberate actions of not-so-ethical people (on purpose), side effects of poor design (engineering mistakes) and, finally, miscellaneous cases related to the impact of the surroundings of the system (environment). Because purposeful design of dangerous AI is just as likely to include all other types of safety problems and will probably have the direst consequences, that is the most dangerous type of AI, and the one most difficult to defend against.

My further research, in collaboration with Federico Pistono (author of "Robots Will Steal Your Job, But That's OK,") explores in depth just how a malevolent AI could be constructed. We also discuss the importance of studying and understanding malicious intelligent software.

Going to the dark side

Cybersecurity research very commonly involves publishing papers about malicious exploits, as well as documenting how to protect cyber-infrastructure. This information exchange between hackers and security experts results in a well-balanced cyber-ecosystem. That balance is not yet present in AI design.

Hundreds of papers have been published on different proposals aimed at creating safe machines. Yet we are the first, to our knowledge, to publish about how to design a malevolent machine. This information, we argue, is of great value – particularly to computer scientists, mathematicians and others who have an interest in AI safety. They are attempting to avoid the spontaneous emergence or the deliberate creation of a dangerous AI.

Whom should we look out for?

Our research allows us to profile potential perpetrators and to anticipate types of attacks. That gives researchers a chance to develop appropriate safety mechanisms. Purposeful creation of malicious AI will likely be attempted by a range of individuals and groups, who will experience varying degrees of competence and success. These include:

- Militaries developing cyber-weapons and robot soldiers to achieve dominance;
- Governments attempting to use AI to establish hegemony, control people, or take down other governments;
- Corporations trying to achieve monopoly, destroying the competition through illegal means;
- Hackers attempting to steal information, resources or destroy cyberinfrastructure targets;
- Doomsday cults attempting to bring the end of the world by any means;
- Psychopaths trying to add their name to history books in any way possible;
- Criminals attempting to develop proxy systems to avoid risk and responsibility;
- AI-risk deniers attempting to support their argument, but making

errors or encountering problems that undermine it;
- Unethical AI safety researchers seeking to justify their funding and secure their jobs by purposefully developing problematic AI.

What might they do?

It would be impossible to provide a complete list of negative outcomes an AI with general reasoning ability would be able to inflict. The situation is even more complicated when considering systems that exceed human capacity. Some potential examples, in order of (subjective) increasing undesirability, are:

- Preventing humans from using resources such as money, land, water, rare elements, organic matter, internet service or computer hardware;
- Subverting the functions of local and federal governments, international corporations, professional societies, and charitable organizations to pursue its own ends, rather than their human-designed purposes;
- Constructing a total surveillance state (or exploitation of an existing one), reducing any notion of privacy to zero – including privacy of thought;
- Enslaving humankind, restricting our freedom to move or otherwise choose what to do with our bodies and minds, as through forced cryonics or concentration camps;
- Abusing and torturing humankind with perfect insight into our physiology to maximize amount of physical or emotional pain, perhaps combining it with a simulated model of us to make the process infinitely long;
- Committing specicide against humankind.

We can expect these sorts of attacks in the future, and perhaps many of them. More worrying is the potential that a superintelligence may be capable of inventing dangers we are not capable of predicting. That makes room for something even worse than we have imagined.

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

Source: The Conversation

Citation: Fighting malevolent AI—artificial intelligence, meet cybersecurity (2016, June 13) retrieved 26 April 2024 from https://phys.org/news/2016-06-malevolent-aiartificial-intelligence-cybersecurity.html