Preventing AI from developing anti-social and potentially harmful behaviour

April 17 2014

Next time you play a computer at chess, think about the implications if you beat it. It could be a very sore loser!

A study just published in the Journal of Experimental & Theoretical Artificial Intelligence reflects upon the growing need for autonomous technology, and suggests that humans should be very careful to prevent future systems from developing anti-social and potentially harmful behaviour.

Modern military and economic pressures require autonomous systems that can react quickly – and without human input. These systems will be required to make rational decisions for themselves.

Researcher Steve Omohundro writes: "When roboticists are asked by nervous onlookers about safety, a common answer is 'We can always unplug it!' But imagine this outcome from the chess robot's point of view. A future in which it is unplugged is a future in which it cannot play or win any games of chess".

Like a plot from The Terminator movie, we are suddenly faced with the prospect of real threat from autonomous systems unless they are designed very carefully. Like a human being or animal seeking self-preservation, a rational machine could exert the following harmful or anti-social behaviours:

- Self-protection, as exampled above.

- Resource acquisition, through cyber theft, manipulation or domination.
- Improved efficiency, through alternative utilisation of resources.
- Self-improvement, such as removing design constraints if doing so is deemed advantageous.

The study highlights the vulnerability of current autonomous systems to hackers and malfunctions, citing past accidents that have caused multi-billion dollars' worth of damage, or loss of human life. Unfortunately, the task of designing more rational systems that can safeguard against the malfunctions that occurred in these accidents is a more complex task that is immediately apparent:

"Harmful systems might at first appear to be harder to design or less powerful than safe systems. Unfortunately, the opposite is the case. Most simple utility functions will cause harmful behaviour and it is easy to design simple utility functions that would be extremely harmful."

This fascinating study concludes by stressing the extreme caution that should be used in designing and deploying future rational technology. It suggests a sequence of provably safe systems should first be developed, and then applied to all future autonomous systems. That should keep future chess robots in check.

More information: "Autonomous technology and the greater human good", by Steve Omohundro, *Journal of Experimental & Theoretical Artificial Intelligence*, published by Taylor & Francis. DOI: 10.1080/0952813X.2014.895111

Provided by Taylor & Francis