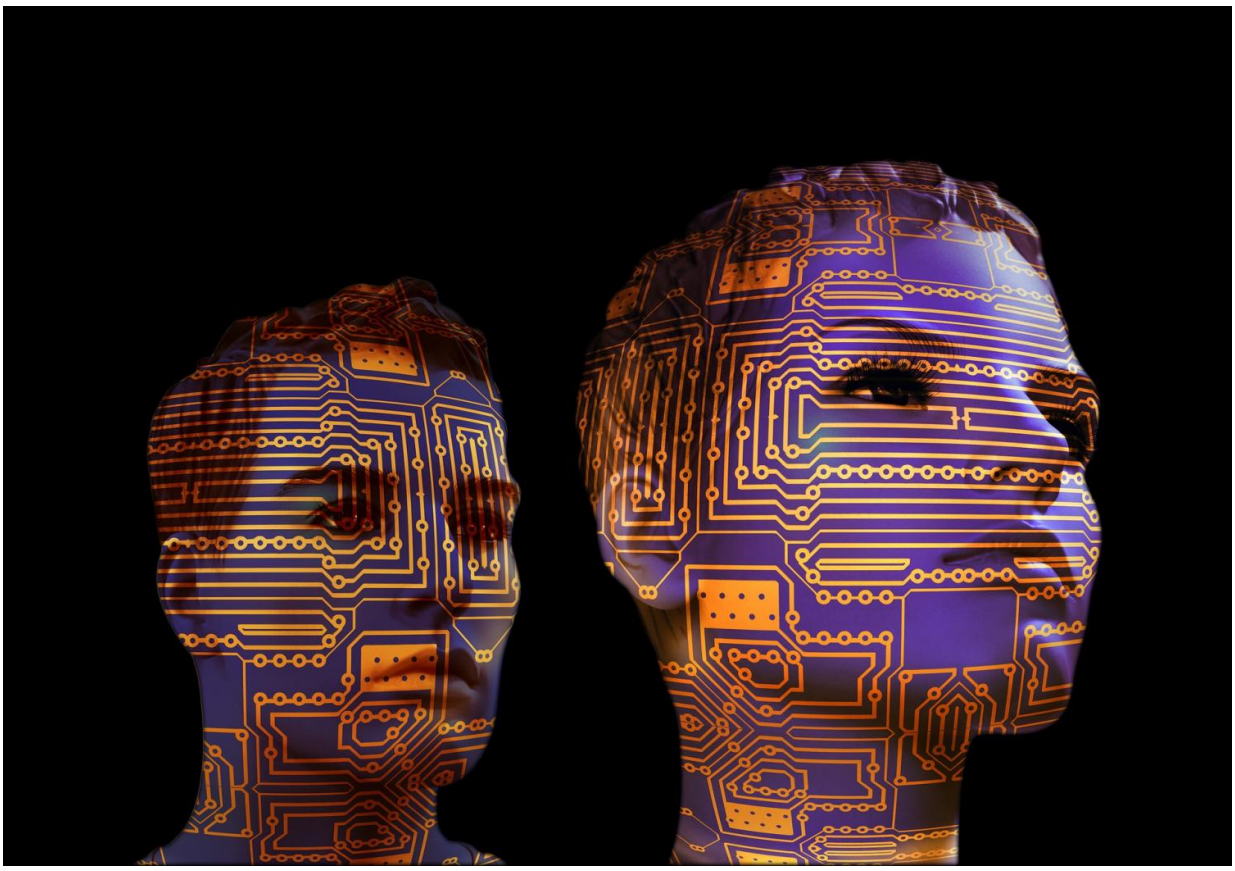
Google team taking upper hand if misbehaving AI agent attempts anything terribly smart

June 6 2016, by Nancy Owano



Credit: CC0 Public Domain

(Tech Xplore)—The twin effects of popular fiction (machines turning

into killer squadrons) and scientific progress (machine learning) have made AI both exciting and scary in its future potential.

Google knows. *TechRadar* and other tech sites are reporting that Google is thinking up a kill switch for dangerous AI.

As *TechRadar* put it, humans can keep the upper hand, for now. David Nield reported: "Google is working on an AI 'kill switch' that allows human operators to turn off super intelligent systems no matter how big their egos get. It's called "safe interruptibility" and it's being developed as part of the DeepMind system."

An open letter went out last year that caused as much of a stir as the worrisome statements preceding it over building superintelligent machines. Notable names in that letter included Elon Musk and Stephen Hawking, and the letter focused on the need to ensure AI research benefits humanity.

"An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence" said that "There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls."

That open letter resonates with the step taken by authors who have released a paper exploring the best ways to prevent self-learning machines from turning off the "big red button" humans might use to shut them down.

Laurent Orseau, Google DeepMind, London, and Stuart Armstrong, The Future of Humanity Institute, University of Oxford, have written "Safely Interruptible Agents."

(DeepMind is the AI research lab owned by Google. DeepMind was founded by Demis Hassabis, Shane Legg and Mustafa Suleyman in London, 2010 and was acquired by Google in early 2014.)

The authors ask, "Given that the human operator has designed a correct reward function for the task, how to make sure that human interventions during the learning process will not induce a bias toward undesirable behaviors?"

They said that "We have proposed a framework to allow a human operator to repeatedly safely interrupt a reinforcement learning agent while making sure the agent will not learn to prevent or induce these interruptions. Safe interruptibility can be useful to take control of a robot that is misbehaving..."

The authors in their abstract pointed out that "now and then it may be necessary for a human operator to press the big red button to prevent the agent from continuing a harmful sequence of actions—harmful either for the agent or for the environment—and lead the agent into a safer situation...This paper explores a way to make sure a learning agent will not learn to prevent (or seek!) being interrupted by the environment or a human operator."

Nield explained what is happening here: "At the heart of the issue is the idea of rewards, or goals that the software code has been programmed to aim for - as AI gets cleverer, it may develop its own rewards that we haven't anticipated (it starts making its own judgements, in other words)."

Timothy Seppala, Associate Editor, *Engadget*, said, "Essentially, DeepMind has developed a framework that'll keep AI from learning how to prevent—or induce—human interruption of whatever it's doing."

If you check out the story postings on this you see numerous references to the red button as the kill switch. The button obviously is not a real hardware piece the authors are constructing; Darren Orf in *Gizmodo* described the nature of the contents: "The paper is filled with math 99 percent of us will never understand, which basically describes methods for building what the paper cheekily calls a 'big red button' into AI."

Tyler Lee in *Ubergizmo* described "a panic button of sorts" to instantly interrupt AI and stop it from doing any harm.

More information: Safely Interruptible Agents (PDF): intelligence.org/files/Interruptibility.pdf