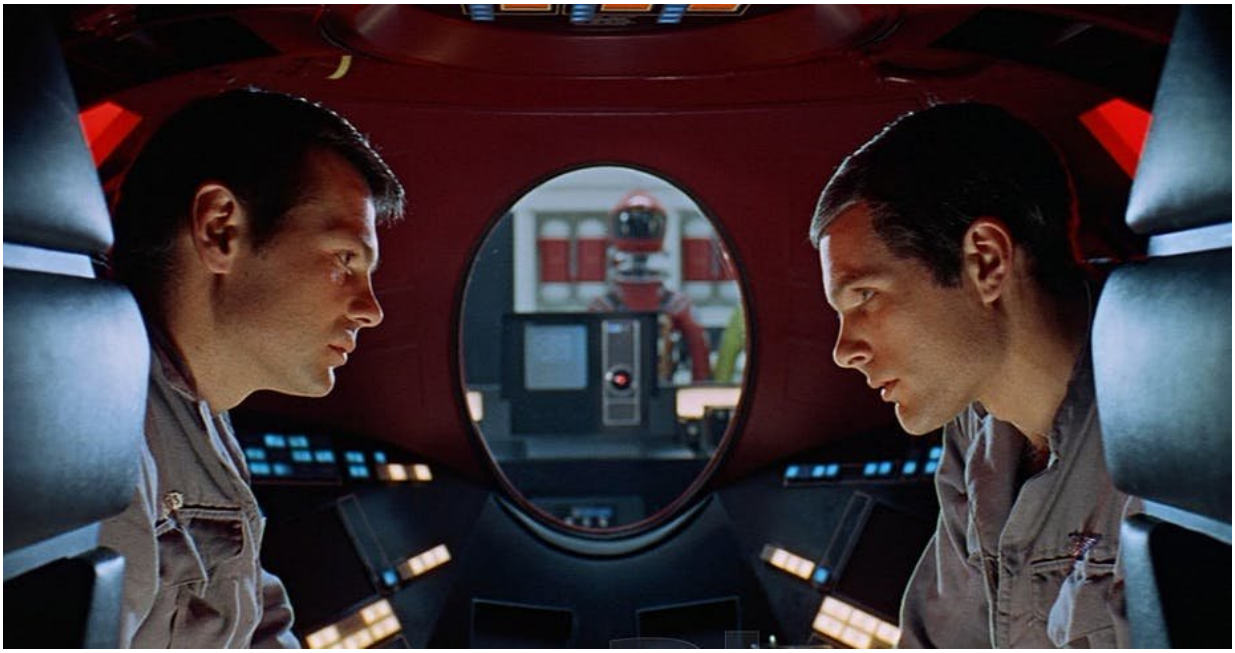
The talking AI story since '2001: A Space Odyssey'

May 10 2018, by Franck Guarnieri



Unfortunately for the intrepid astronauts in Stanley Kubrick's 2001, a Space Odyssey, the HAL 9000 computer knows how to read lips. Credit: IMDB

Almost everyone knows the story of HAL 9000, the killer supercomputer in Stanley Kubrick's landmark film *2001: A Space Odyssey*, whose 50th anniversary will be celebrated on May 12, 2018 at the 71st Cannes Film Festival. In an intriguing scheduling coincidence, IBM, Kubrick's partner during the filming of *A Space Odyssey*, and Airbus have just unveiled the CIMON (Crew Interactive Mobile

Companion) project, an "intelligent, mobile and interactive astronaut assistance system" that will join the International Space Station.

These two events propel us into a debate over the risks created by the development of superintelligence that could eliminate jobs on a massive scale or, even worse, wipe the human species off the face of the planet – and raise the question of how to assess such a threat.

To date, we have no experience of accidents or disasters due to faulty or malicious AI. However, the imaginations of artists and scientists are a treasure trove of material that tells the story of superintelligence freed from any human control.

A story of AI going wrong

2001: A Space Odyssey is a forerunner of contemporary controversies. It tells the story of the struggle and eventual conquest of a human being, the only survivor of a methodical programme of extermination led by a sentient supercomputer.

Aboard the spaceship *Discovery One*, only the supercomputer HAL 9000 has been informed by its creators of the purpose of the mission: to reach Jupiter and search for signs of extra-terrestrial intelligence. Although considered to be infallible, HAL makes an error. The machine refuses to admit this, and, caught out, it claims that the mistake is due to "human error". In principle the humans are the computer's designers but, if it is to be believed, could it in fact be the computer itself? Adopting this line of reasoning, the machine gives itself a status that crew members could not imagine – that of a living, sentient and thinking being.

For the crew, HAL's error is unacceptable. There is no room for forgiveness or charity: error may be human, but it is not machine. There is no appeal: HAL must be taken out of service. The supercomputer,

omnipresent and omniscient, immediately discovers the project designed to end its life. To survive and complete its mission, it decides to eliminate the crew. Only one human survives, Astronaut David Bowman, and he resumes, with even more determination, the digital homicide mission.

Bowman succeeds in penetrating the core of the unit and then mechanically, emotionlessly and almost ceremoniously disconnects, one by one, the machine's memory circuits from their housing. Like a child caught with its hand in the cookie jar, the computer tries, by talking about itself, to derail the lobotomy. In a final attempt, it sings a song it learned in its first hours of "life", but nothing works, and finally its voice fades away.

AI that can talk about itself

Much more than the story of a fight to the death, one of the treasures of the film is that is considers the narrative dimension of AI. It's able not only to make up stories about itself, but also can fail due to circumstances and its own errors. Thus, the elimination of the crew is not the result of HAL 9000 becoming autonomous, but from a "bad story" that the machine tells itself, that of believing that the crew could compromise the mission.

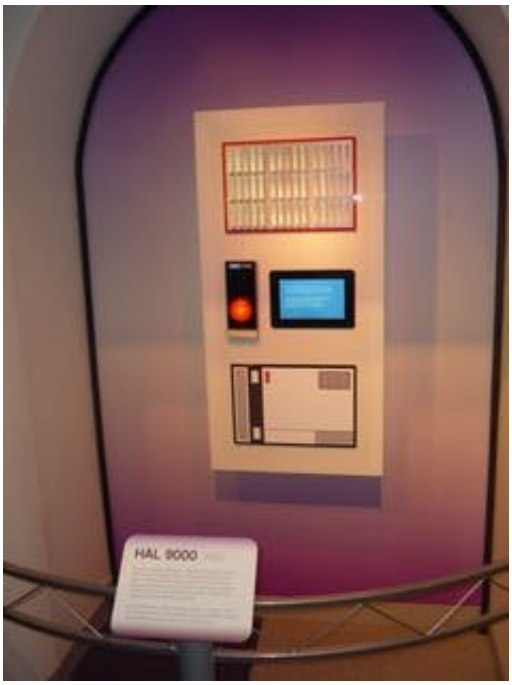
Kubrick's work thus makes it possible to conceive the risks caused by superintelligence not in terms of technical domination, but as the construction of an imperfect narrative identity. Although reality is still far from catching up with fiction, some initial findings in this area are worth thinking about.

In 2016, a novel titled *The Day a Computer Writes a Novel… Almost* won a Japanese literary prize, the Nikkei Hoshi Shinichi. It was prewritten by an AI research team from the University of Hakodate, whose work

initially consisted of selecting words and sentences, then defining parameters that allowed a program to "write" the novel. Of more than 1,450 submissions for the prize, 11 had been written, at least in part, by a non-human. Of course, the jury knew nothing about the "authors".

The following year, in the same spirit but without a digital co-author, Zack Thoutt, a fan of the TV series *Game of Thrones*, used a neural network fed with over 5,000 pages of text from the books on which it was based to predict what might happen next in the story. The result is far from equal to the work of author George R. Martin, but the text is, for the most part, understandable and the predictions are similar to some of the theories that are popular with fans of the series.

More recently, a program designed by MIT's Media Lab – baptised "Shelley" after Mary Shelley, the author of *Frankenstein; or, The Modern Prometheus* – created terrifying tales, designed to scare. The first stage was to train the program with stories written by humans taken from a database of over 140,000 references. In the second stage, the computer generated its own works that it improved by collaborating with humans who responded to its messages via Twitter.



HAL 9000 au Robot Hall of Fame. Credit: Photojunkie/Wikipédia, CC BY

And in the domain of reading and understanding stories, Google has just launched its "Talk to book" service that allows users to converse in natural language with an automatic learning algorithm that is supposed to help them in their future reading choices.

Stories that evoke the risks of AI

If, like HAL 9000, AI computers try to write stories, they are still far from reaching the standards set by human authors. Although no one can predict with certainty what artificial superintelligence might look like, it remains possible to imagine it by producing stories that stimulate our thinking.

The historian Yuval Noah Harari undertook the task, imagining a future where the automation of machines would cause the disappearance of the

majority of jobs. He argued that humans risk losing their economic value because intelligence will be decoupled from consciousness. While intelligence is necessary to drive a car or diagnose a disease, consciousness and subjective human experiences are not mandatory.

Even more alarmist, the philosopher Nick Bostrom puts forward the idea that humans will probably not have the opportunity to experience this disastrous revolution because it is likely they will be exterminated as soon as artificial superintelligence appears.

Both Harari and Bostrom base their conclusions on a reduction of action to its functional dimension, judged solely in terms of effectiveness. But such a vision neglects whole areas of human existence. To remedy this, neuroscientist Antonio Damasio argues that life represents a complex act in which feelings are the expression of a permanent struggle to achieve a balance that underpins human existence. For Damasio, without subjectivity there is no creativity, and without the emotions that are manifest in the relationships between the body and the brain to perceive reality, there is no humanity.

Accepting this relegates Kubrick to the rank of a genius whose prophecy, albeit poetic, is unrealistic. It also means admitting that an AI computer will never be able to redefine its own mission at the expense of its creators. And finally, we must take very seriously the idea that the main risk to humanity is a cyberwar initiated by humans themselves – a conflict populated by drones and killer robots supported by swarms of computer viruses that all smarter than the others.

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

Provided by The Conversation