**PRESS RELEASE**
**NOTA DE PRENSA**
28 MAY 2018, MADRID

Publication date: 23 August 2016, Madrid

# Containing a Superintelligent AI Is Theoretically Impossible

**Source(s):** Motherboard

We reproduce here in full an article published by the online magazine and video channel Motherboard of Vice Media. The article is based on the publication 'Superintelligence cannot be contained: Lessons from Computability Theory', co-authored by Antonio Fernández Anta, Research Professor at IMDEA Networks Institute.

---

Machines that "learn" and make decisions on their own are proliferating in our daily lives via social networks and smartphones, and experts are already thinking about how we can engineer them so that they don't go rogue.

So far, suggestions have ranged from "training" self-learning machines to ignore certain kinds of information that might teach them racism or sexism, to coding them with values like empathy and respect. But according to some new work from researchers at the Universidad Autónoma de Madrid,as well as other schools in Spain, the US, and Australia, once an AI becomes "superintelligent"—think *Ex Machina*—it will be impossible to contain it.

Well, the researchers use the word "incomputable" in their paper, posted on the ArXiv preprint server, which in the world of theoretical computer science is perhaps even more damning. The crux of the matter is the "halting problem" devised by Alan Turing, which holds that no algorithm is able to correctly predict whether another algorithm will run forever or whether it will eventually halt—that is, stop running.

Imagine a superintelligent AI with a program that contains every other program in existence. The researchers provided a logical proof that if such an AI could be contained, then the halting problem would by definition be solved. To contain that AI, the argument is that you'd have to simulate it first, but it already simulates everything else, and so we arrive at a paradox.

**"It would not be feasible to make sure that [an AI] won't ever cause harm to humans"**

"This is a standard methodology in theoretical computer science," Manuel Alfonseca, a professor at the Universidad Autónoma de Madrid'sEscuela Politécnica Superior and lead author of the work, wrote me in an email. "You prove that a problem can be reduced to another one, and then what you know about the second problem can be applied to the first."

Basically, since the halting problem is, in theoretical computer science lingo, "undecidable," then containment of an AI is computationally impossible.

Of course, the very idea of a superintelligent AI capable of simulating the entire state of the world is thought by some to be merely theoretical. Despite warnings from technology luminaries and philosophers, many computer scientists believe that superintelligence is more science fiction than

science.

The halting problem has been used to think through other tricky issues relating to the future of advanced artificial intelligence. In 2014, a team of researchers concluded that a similar problem existed for the ethics of machines: it is theoretically impossible to design an algorithm that can correctly predict whether another algorithm will act “morally.”

But, again, this is all theory, not a guarantee about the future material state of the world. The solution to the moral aspect of the halting problem might be, as Motherboard editor Michael Byrne wrote, to prove with a degree of certainty through tests that a given AI will act morally. After all, just because one algorithm can’t predict the behaviour of another, doesn’t mean that we can’t make the *first* algorithm behave properly to begin with.

Alfonseca isn’t so sure about a similar solution to the containment problem.

“I believe that if [superintelligent] AI were possible, it would not be feasible to make sure that it won't ever cause harm to humans,” he wrote.

Let’s just hope we never get there, then.

*Written by Jordan Pearson, Motherboard Staff Writer (Canada) – 8 July 2016*

***Image: Flickr/Victory of the People***

**Bibliographical references:**

‘Superintelligence cannot be contained: Lessons from Computability Theory’ [PDF 1.4 MB]
Manuel Alfonseca[1], Manuel Cebrian[2], Antonio Fernandez Anta[3], Lorenzo Coviello[4], Andres Abeliuk[5], and Iyad Rahwan[6]
*[1] Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain [2] Data61 Unit, Commonwealth Scientific and Industrial Research Organisation, Melbourne, Victoria, Australia*
*[3] IMDEA Networks Institute, Madrid, Spain*
*[4] Google, USA*
*[5] Melbourne School of Engineering, University of Melbourne, Melbourne, Australia*
*[6] The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, US*

###

Traducción al español:
/noticias/2016/contener-una-inteligencia-artificial-ia-superinteligente-teoreticamente-imposible
Original source:
news/2016/containing-superintelligent-ai-theoretically-impossible

## About Us

**IMDEA Networks Institute** is a **research organization on computer and communication networks** whose multinational team is engaged in cutting-edge fundamental science and technology. As a growing, English-speaking institute located in Madrid, Spain, IMDEA Networks offers a unique opportunity for pioneering scientists to develop their ideas. IMDEA Networks has established itself internationally at the forefront in the **development of future network principles and technologies.** Our **team** of highly-reputed researchers is designing and creating today the networks of tomorrow.

Read more on www.networks.imdea.org.

***Some keywords that define us:*** *5G, Big Data, blockchains and distributed ledgers, cloud computing, content-delivery networks, data analytics, energy-efficient networks, fog and edge computing, indoor positioning, Internet of Things (IoT), machine learning, millimeter-wave communication, mobile computing, network economics, network measurements, network security, networked systems, network protocols and algorithms, network virtualization (software defined networks – SDN and network function virtualization – NFV), privacy, social networks, underwater networks, vehicular networks, wireless networks and more...*

IMDEA Networks Institute
Avda. del Mar Mediterráneo, 22
28918 Leganes (Madrid) Spain
@IMDEA_Networks | Linkedin | Facebook

**Telephone:** +34 91 481 6210
**E-mail:** mediarelations.networks@imdea.org
**Web**: www.networks.imdea.org