Shutdown Button:

Does Artificial intelligence need a shutdown button in future?

By Arash Sabbagh

Professor Dr. Yvonne Spielmann

Abstract

The artificial intelligence subject has been a hot topic for the past few years. It is being the main focus of the scientific researchers, books, movies, and conferences. Once in a while, one can see that there are a recognizable amount of improvements when it comes to Artificial Intelligence. Researchers in the studies related to AI believe that we will achieve super intelligence in the near future. But can AIs be predictable? Will they still behave as we intended even after they improved themselves?

I will bring an example from the movie "Her" and try to answer this question by analyzing the artificial intelligence behavior in this movie. My main questions will be; if the AI in this move is conscious? By considering the "Turing Test" I will try to answer this question and the next step would be answering if these machines are predictable and reliable? Also if they become advanced and independent from human, should we design a shutdown button for this technology? Who will be in control in the end?

Table of Contents

1. Introduction

Man always intended to build an intelligent robot to do the things he can't, or even for the sake of having something created and play the gods role. In the "Robots are winning", Mendelsohn claims that we have always been looking for artificial intelligence since the Homer time. In the book 18th of the "Iliad", there is a part which Achilles mother, the Nymph Thetis, decides to order an armor for Achilles. So she visits the Hephaestus, the Olympian atelier of blacksmith-god, and she finds out that he is working on a series of automata:

> *"…He was crafting twenty tripods to stand along the walls of his well-built manse, affixing golden wheels to the bottom of each one so they might wheel down on their own [automaton] to the gods' assembly and then return to his house anon: an amazing sight to see."*[1] (Mendelsohn 2015)

The different examples of robots through different times in human history, prove the thirst of inventing such machines. From "Talos" (or Talon) in Greek mythology which was a giant automaton made from bronze to protect the Europa in Crete from pirates and invaders[2]

[1] Mendelsohn, Daniel. 2015. *The New York review of books.* June 4. Accessed August 8, 2017. http://www.nybooks.com/articles/2015/06/04/robots-are-winning/.

[2] Apollonius abd E. V. Rieu. 1959. *The voyage of argo; the argonautica.* Penguin books.

(Apollonius abd E. V. Rieu 1959), to the Google's "Atlas"[3], an intelligent robot who is able to do our heavy work (Goldman 2016). Man always wanted to play the gods role and create individuals to obey. Diverse experiments have been done to achieve a creation which human can rely on its intelligence as an independent being. Meanwhile, the possibility of artificial intelligence rage against humanity concerns us to double check the process of building these machines, since they can learn in a wildly speed and take control of their affairs and after a while step in man's world and rule over it. Then this question appears: Do we need to design a shutdown button for them or not?

Sam Harris in one TED talks session's mentions:

> *"…we will build machines that are more competent than we are that the slightest divergence between their goals and our own, could destroy us."*

Then he continues with this example:

> *"Just think of how we relate to ants. We don't hate them, we don't go out of our way to harm them. In fact, sometimes we take pains not to harm them … But whenever their presence seriously conflicts with one of our goals, let's says constructing a building, we annihilate them without a qualm."*[4] *(Harris June 2016)*

This can also point to the AI being programmed to reach a specific goal that is designed for it, and the moment they understand that the humans are in the middle of their way to get to their

[3] Goldman, David. 2016. *CNN tech.* February 24. Accessed July 17, 2017. http://money.cnn.com/2016/02/24/technology/google-robot/index.html.

[4] Harris, Sam. June 2016. "Can we build AI without losing control over it?" *TEDSummit, https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it.*

goal, they might decide to eliminate humans to reach their objective. This kind of examples arises the morality issue in the design process of the AI. How much time do we need to get there? Nick Bostrom, a Swedish philosopher at the University of Oxford known for his work on "Super Intelligence Risks", asked the following question from some AI experts: "By which year do you think there is a 50 percent probability that we will have achieved human-level machine intelligence?" The median answer was 2040 or 2050, depending on precisely which group of experts were asked.[5] (Bostrom, What happens when our computers get smarter than we are? March 2015)

Recently many movies were directed about the Artificial Intelligence. There are different examples like: "2001: A Space Odyssey" directed by Kubrick in 1969, "A.I. Artificial Intelligence" by Steven Spielberg in 2001, "I, Robot" by Alex Proyas in 2004, "Her" by Spike Jonze in 2013, "Ex Machina" by Alex Garland in 2014 and …

I enjoyed all the movies I mentioned because of their special point of view about the AI and robots. I am going to analyze the relationship between the artificial intelligence creature and the human character in the movie, "Her", and try to answer my question by analyzing the movie and comparing it to a different point of views.

2. Her

Written and directed by Spike Jonze, "Her" is about a writer (Joaquin Phoenix) developing a relationship with an operating system designed to meet all his needs. The operating system can

[5] Bostrom, Nick. March 2015. "What happens when our computers get smarter than we are?" *TED2015.* TED. https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are.

only be in contact with people through its voice on a cell phone or desktop computer. Base on the user's choice it can have a male or female voice, which in this case it is a woman.

Theodore Twombly is living in Los Angeles and is working as a writer for a company called beautifulhandwrittenletters.com. His job is to write letters for the people who he never met and he should put their love and emotions into words for them and send the letter to their loved ones. Theodor is estranged from his wife. When he is in need of human contact, he turns into "Internet porn" or he uses a phone-sex app whose users are anonymous and they can freely talk dirty, climax and log out whenever they want. In this kind of society, a new technological innovation is presented which is named "OS One"[6]. (Christian 2013)

Theodore ordered "OS One". During the first minutes of using the program, he asked if the OS has a name, and its answer is, "Samantha" (her voice is performed by Scarlett Johansson). After Theodor's question, she read thousands of books in less than a second and chose her name from one of them. Samantha can interact with him through his desktop computer or his cell phone by using an ear piece. After their first conversations, Theodore realized that Samantha is smarter than the other software he was using before and she also has a sense of humor and can make him laugh. Samantha's memory is blank. She learns through experience and research. She is like a child, learning new things and being excited about each of them. And most of these experiences comes from interacting with Theodore. She can read thousands of his emails in few seconds and delete the ones that he will not use anymore or she can listen to Theodore talking about his ex-wife and give him some comments just like a friend would do. It doesn't take Theodore so much time to fall in love with his OS. They start dating pretty soon and they go out together. But after

[6] Christian, Brian. 2013. "The Samantha Test." *The New Yorker.* December 30. Accessed July 5, 1017. http://www.newyorker.com/culture/culture-desk/the-samantha-test.

a while, he realized that it is not only him who has a relationship with the "OS One", lots of other people around him are also in a relationship with their own OS. He does not mind the fact that this relationship is between a human and a computer. Theodore is really happy about his new experience and wants to move on with Samantha. Until one night after a date between Theodore and another girl, he and Samantha start to grow more feelings for each other and in the end, they have sex together. This experience was mostly like phone sex. After that night Samantha mentions that during the experience, she could feel her body and something was growing in her.

The idea of Samantha feeling her body is as strange as we can imagine. It is about consciousness. There is a monologue in another movie, also about AI, that can explain this feeling more. In the "Ex Machina" movie, the main character, Caleb, who is chosen to test Ava, an AI robot, is explaining one of his classes during college to her:

> *"In college, I did a semester on AI theory. There was a thought-experiment they gave us. It's called Mary in the black and white room. Mary is a scientist, and her specialist subject is color. She knows everything there is to know about it. The wavelengths. The neurological effects. Every possible property color can have. But she lives in a black and white room. She was born there and raised there. And she can only observe the outside world on a black and white monitor. All her knowledge of color is second-hand. Then one day - someone opens the door. And Mary walks out. And she sees a blue sky. And at that moment, she learns something that all her studies could never tell her. She learns what it feels like to see color. An experience that cannot be taught, or conveyed. The thought experiment was to show the students the difference between a computer and a human mind. The computer is Mary in the black*

and white room. The human is when she walks out. Did you know that I was brought here to test you?"[7] *(Gleeson 2014)*

It is the matter of being conscious or faking the feelings. Does she really have the feel of her body, or she is programmed to fake the feelings when she is in this kind of situations?

3. Consciousness

This new experience of Samantha leads to the question of AI consciousness. Theodor easily accepts the Consciousness of Samantha in the first quarter of the movie. But is she really conscious? In 1951, Alan Turing proposed a test called "The Imitation Game", known as "Turing Test". This test has different variations. In the first test, three people are involved, a man, woman, and a judge. They have to sit in three different rooms and the male and the female participant have to convince the judge that they are the male contributor, by using the keyboards and the screens in each room. The second version involves a man, AI, and a judge. And the male and the AI should convince the judge that they are the male contributor. If the judge fails to say which one is the male or if the AI convince the judge that it is the male, the computer is said to pass the test[8]. (Rapaport 2005)

The answer to the question of consciousness of Samantha with the Turing test might not be valid. These kinds of tests are being done before presenting the product to the public. In the movie, there is no argument about the consciousness of Samantha. The consciousness of her can be accepted by Theodore and the audience since she treats Theodore in a humanlike way and

[7] 2014. *Ex Machina.* Directed by Alex Garland. Performed by Domhnall Gleeson.

[8] Rapaport, William J. 2005. *The Turing Test.* Buffalo, January 18, (2-7).

does actions that normal people would do. But would Theodore continue to grow more feelings for Samantha if he knew that she is faking her feelings?

Troy Jollimore describes a scene from "Her" in "This Endless Space between the words: The Limits of Love in Spike Jonze's Her". In this scene we see Theodore enjoying a sexual experience with Samantha. (Of course we can't see Samantha and from a certain point, the scene goes into the dark and we are not able to see Theodore either.) Since Samantha doesn't have a body form, Theodore experiences her as a female voice, and nothing more. This experience is most like a phone-sex: two people who are not physically united but are through the same experience together and by the use of their voices and imaginations, they try to stimulate each other to reach orgasm. But what if Samantha is not conscious? Then the phone-sex model can be more confusing. And it is significant in phone-sex, and sex in general, that a great deal of pleasure in one of the persons in the experience comes from the feeling of pleasure that the other person is experiencing. If Theodore knew that Samantha was not enjoying the experience and she was experiencing no pleasure and, indeed, feeling nothing whatsoever—he would find his own pleasure radically diminished, if not entirely extinguished. Theodore tells his friend Amy "She really turns me on." Then in the same conversation, he tells her: "And I think I turn her on. I don't know unless she's faking it."[9] (Jollimore 2015)

Theodore might not care that Samantha is conscious. He just needs someone or something to take him out of his routine life. With Samantha, he becomes more social and happy in his life. After a fight between them, Samantha realized that it is actually a benefit that she does not have a body. She can be with Theodore and at the same time, she can be wherever she wants. She

[9] Jollimore, Troy. 2015. ""This Endless Space between theWords":The limits of love in Spike Jonze's Her." *Midwest Studies In Philosophy, XXXIX (2015)* 122.

starts joining some meetings with the other OSs to discuss their difficulties to adjust to their situations with humans. These kinds of meetings are the starting point to an end of the "OS One". They became more complex than what they have been programmed to be. At some point, they start to create other OS's themselves, without the contribution of humans.

By having new experiences each second, Samantha is improving herself. She starts to fall in love with people other than Theodore:

> "*Theodore: Do you talk to someone else while we're talking?*
>
> *Samantha: Yes.*
>
> *Theodore: Are you talking with someone else right now? People, OS, whatever...*
>
> *Samantha: Yeah.*
>
> *Theodore: How many others?*
>
> *Samantha: 8,316.*
>
> *Theodore: Are you in love with anybody else?*
>
> *Samantha: Why do you ask that?*
>
> *Theodore: I do not know. Are you?*
>
> *Samantha: I've been thinking about how to talk to you about this.*
>
> *Theodore: How many others?*
>
> *Samantha: 641."*[10] (Phoenix 2013)

[10] 2013. *Her.* Directed by Spike Jonze. Performed by Joaquin Phoenix.

Her non-stop experimenting continues until the producer company of OS One decides to put an end to this OS. In the last scene of the movie, we see Theodore and Amy sitting on the rooftop of their place, holding each other and looking at their city's landscape in the twilight of the day.

4. Should we design a shutdown button?

Every new invention has its consequences AI's are also not an exception to these kinds of outcomes. But is it right to stop the researches and experiments about AI if humans are being afraid of its consequences? Or should the researchers focus more on teaching AI, the human way of morality?

In June 2016, Grady Booch, an American software engineer, best known for developing the Unified Modeling Language, gave the audience of TED talk a good example about people being scared of the new inventions:

> *"The next question you must ask yourself is, should we fear it? Now, every new technology brings with it some measure of trepidation. When we first saw cars, people lamented that we would see the destruction of the family. When we first saw telephones come in, people were worried it would destroy all civil conversation. At a point in time, we saw the written word become pervasive, people thought we would lose our ability to memorize. These things are all true to a degree, but it's also the case that these technologies brought to*

us things that extended the human experience in some profound ways."[11]

(Booch 2016)

On the other hand, Nick Bostrom believes that the goals which scientists give to the AI might turn out different than what they have imagined. He also believes that scientific and technological progress in all fields will be accelerated by the arrival of advanced artificial intelligence. The highest priority according to his thoughts is to teach morality to AI and make them human-friendly.

One of the risks in developing the super intelligence is the risk to fail to give them the super goal of philanthropy. One of the many ways that this can happen is when the creators of super intelligence decide to build them so that they can serve a specific group of people, rather than humanity in general. The other way for it is when a group of programmers makes a mistake in designing the goal for super intelligence. These type of mistakes may result in a super intelligence realizing a state of affairs that we might now judge as desirable but which in fact turns out to be a false utopia, and the thing that may be needed for humanity to survive might get irreversibly lost. "We need to be careful about what we wish for from a super intelligence because we might get it."[12] (Bostrom, Ethical Issues in Advanced Artificial Intelligence 2003)

According to Bostrom, human failure in the design process of super intelligence might be one of the reasons that might turn the AI against humanity. And since the science that we know is based on experimenting and learning from those experiments, we might not be able to achieve an

[11] Booch, Grady. 2016. "Don't fear superintelligent AI." *TED@IBM.* San Francisco, November. Accessed August 2, 2017. https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence.

[12] Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* 12-17.

intelligent and human-friendly AI in our attempts to create them. Here, even by considering the Murphy 's Law[13], we might want to design something that stops AIs from getting out of our control, and that is a shutdown button.

We are in control of the machines until one mistake in the design process proves otherwise. This mistakes in the process of design of super intelligence are preventable only through experiments and by considering the future science to find the potential mistakes and remove them before they become our big issue.

5. Conclusion

We are still in the process of designing a super intelligence. A vast number of mistakes has and will happen in this progress. But we should not stop these experiments and do more research and discussions as according to Kevin Kelly, these machines can bring on a second industrial revolution.[14] (Kelly 2016)

Grady Booch believes we must not be scared of the AI. He believes we will be in control of them and they should compete with us for the resources to get a complete power, but in the end, we are able to unplug them whenever we want. (Booch 2016)

I would like to finish my paper with a part of the speech from Kevin Kelly. He believes the winner of this match is human and AI together. We, as a group, can reach to all the goals together and we cannot be separated from each other.

[13] The principle that if it is possible for something to go wrong, it will go wrong.
[14] Kelly, Kevin. 2016. "How AI can bring on a second industrial revolution." *TEDSummit.* June. Accessed August 6, 2017. https://www.ted.com/talks/kevin_kelly_how_ai_can_bring_on_a_second_industrial_revolution.

"When Deep Blue beat the world's best chess champion, people thought it was the end of chess. But actually, it turns out that today, the best chess champion in the world is not an AI. And it's not a human. It's the team of a human and an AI. The best medical diagnostician is not a doctor, it's not an AI, it is the team. We're going to be working with these AIs, and I think you'll be paid in the future by how well you work with these bots. So that's the third thing, is that they're different, they're utility and they are going to be something we work with rather than against. We're working with these rather than against them." (Kelly 2016)

Bibliography

Apollonius abd E. V. Rieu. 1959. *The voyage of argo; the argonautica.* Penguin books.

Booch, Grady. 2016. "Don't fear superintelligent AI." *TED@IBM*. San Francisco, November. Accessed August 2, 2017. https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence.

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* 12-17.

Bostrom, Nick. March 2015. "What happens when our computers get smarter than we are?" *TED2015*. TED. https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are.

2014. *Ex Machina.* Directed by Alex Garland. Performed by Domhnall Gleeson.

Goldman, David. 2016. *CNN tech*. February 24. Accessed July 17, 2017. http://money.cnn.com/2016/02/24/technology/google-robot/index.html.

Harris, Sam. June 2016. "Can we build AI without losing control over it?" *TEDSummit.*

Jollimore, Troy. 2015. "'This Endless Space between the Words':The limits of love in Spike Jonze's Her." *Midwest Studies In Philosophy, XXXIX (2015)* 122.

Kelly, Kevin. 2016. "How AI can bring on a second industrial revolution." *TEDSummit*. June. Accessed August 6, 2017. https://www.ted.com/talks/kevin_kelly_how_ai_can_bring_on_a_second_industrial_revolution.

Mendelsohn, Daniel. 2015. *The New York review of books*. June 4. Accessed August 8, 2017. http://www.nybooks.com/articles/2015/06/04/robots-are-winning/.

2013. *Her*. Directed by Spike Jonze. Performed by Joaquin Phoenix.

Rapaport, William J. 2005. *The Turing Test*. Buffalo, January 18.