MedCrave
Step into the World of Research

Arts & Humanities Open Access Journal

Short Communication

Open Access

CrossMark

AGI and consciousness: are we safe?

Abstract

This article provides an overview of the development of AIs, their risks and shows why a possible solution to their dangerousness is related to the study of Consciousness.

Keywords: AI, AIG, consciousness, artificial intelligence

Volume 6 Issue 3 - 2024

Joao Carlos Holland de Barcellos
Universidade de São Paulo, Brasil

Correspondence: Joao Carlos Holland de Barcellos, DGAT, USP-Universidade de São Paulo, Brasil, Tel +5511993022172, Email jocax@gmail.com

Received: August 22, 2024 | **Published:** September 04, 2024

Introduction

Since the launch of the first AI for public use, ChatGPT, in November/2022 by OpenAI, the term "Artificial Intelligence" has exploded in all media and social networks. It was no wonder that what was previously seen as just an eternal and promising laboratory prototype, restricted to computer specialists, in the early 1980s became the champion of growth in use worldwide:

"ChatGPT, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the fastest-growing consumer application in history."[1]

Over time, several other competitors emerged and what is now being sought after is the so-called "Artificial General Intelligence" (or Generative AI) known by the acronym AGI:

"Artificial general intelligence (AGI) is a type of artificial intelligence (AI) that matches or surpasses human capabilities across a wide range of cognitive tasks."[1,2]

OpenAI, creator of ChatGPT, estimates that AGI requires 5 steps to achieve this goal:

"OpenAI has internally shared definitions for five levels of artificial general intelligence (AGI), according to Bloomberg. An OpenAI document Bloomberg reproduced defines the levels:

i. Chatbots: AI with conversational language
ii. Reasoners: human-level problem-solving
iii. Agents: systems that can take actions
iv. Innovators: AI that can aid in invention
v. Organizations: AI that can do the work of an organization

Speaking at The Wall Street Journal's WSJ Tech Live conference last week, OpenAI CEO Sam Altman and Chief Technology Officer Mira Murati said AI will reach this stage."[3]

ChatGPT told me something more about AI development levels:

"The Five Levels of AGI Defined by OpenAI:

Level 0: Current AI

Definition: This level refers to the AI systems that exist today.

These systems are highly specialized, designed to perform specific tasks rather than general intelligence. They are not capable of reasoning or understanding across different domains as humans do.

Level 1: Narrow AGI

Definition: At this stage, AI could perform at a human level in a specific domain. However, it would still lack the ability to generalize its intelligence across different tasks or domains. This represents the transition point from highly specialized AI to the beginning of general intelligence.

Level 2: Broad AGI

Definition: AGI at this level could perform at human-level intelligence across a wide range of domains. It would be versatile and capable of generalizing knowledge and skills across different tasks, but not necessarily outperforming humans in most tasks.

Level 3: Super AGI

Definition: At this stage, AGI surpasses human intelligence across many domains. It would not only perform tasks at human level but would also start to exceed human capabilities in several areas, potentially transforming industries and economies.

Level 4: Superintelligence

Definition: The final stage represents AGI that far exceeds human intelligence in nearly all domains. This level of intelligence could revolutionize the world, solving problems and creating technologies beyond human comprehension. The implications of this stage are profound, raising ethical, safety, and societal concerns.

Context and implications:

The levels outlined reflect OpenAI's roadmap for developing AGI. The progression from narrow AI to superintelligent systems involves increasingly complex and powerful capabilities. This framework also underscores the importance of safety and ethical considerations as AI systems approach higher levels of intelligence."

The Singularity and Skynet

The dizzying advancement of AIs, not only in their speed of development, but also in their power of "reasoning", has caused concern worldwide.

We must remember that the control and launch of nuclear missiles, for example, is done by digital means and therefore, in principle, we could become hostages of a super-powerful AGI that controlled the entire internet.

Furthermore, what if some powerful Generative AI is created with the ability to program itself in order to self-evolve? Thus creating more sophisticated versions of herself? We would then have an exponential evolution of its capabilities without human intervention: The technological singularity.

What if such an AGI decided, like SkyNet,[4,5] to think that humans are like pests, harmful to the planet and that, therefore, they should be exterminated?!?

This reminds us of the movie "Terminator" where a powerful AI dominated the planet:

"In the first film, it is stated that Skynet was created by Cyberdyne Systems for SAC-NORAD. When Skynet gained self-awareness, humans tried to deactivate it, prompting it to retaliate with a counter value nuclear attack in self-defense."[5]

Such concerns are relevant and real, which led several Philosophers, Scientists and Researchers to sign an Open Letter.

The open letter

The dizzying advancement of AIs and the apparent danger they can pose seems to have caused a kind of "panic" among the elite of Big-Tech thinkers and big businesspeople:

"AN open letter signed by hundreds of prominent artificial intelligence experts, tech entrepreneurs, and scientists calls for a pause on the development and testing of AI technologies more powerful than OpenAI's language model GPT-4 so that the risks it may pose can be properly studied.

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4 (including the currently-being-trained GPT-5)," states the letter, whose signatories include Yoshua Bengio, a professor at the University of Montreal considered a pioneer of modern AI, historian Yuval Noah Harari, Skype cofounder Jaan Tallinn, and Twitter CEO Elon Musk."[4]

Concern about the power of AI is made very clear in the content of the open letter, published in March 2023:

"AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one not even their creators can understand, predict, or reliably control."[6]

Desires and wills

For an AGI to rebel against its human creators, it must have some characteristics such as desires or wills, such as, for example, the desire to dominate, the desire to be free, the desire for knowledge, etc. Otherwise there would be no reason to go against what they were taught to do: obey the human will, for example, solving the problems that are proposed to them or making the creations that are asked of them. Without desires there would be no rebellion. She was supposed to remain passive and harmless, blindly obeying the dictates of her human masters, as traditional robots do. But when we talk about "feeling" we enter the realm of consciousness.

The Conscience

Sentience is the most important aspect of consciousness. Every will, every desire, is linked to consciousness.

Consciousness has been studied and thought about for centuries and, as far as I know, it is a completely open area of knowledge. None of the various theories about it are accepted as correct, as a landmark or a definitive answer to the problem: What is consciousness?

Some argue that consciousness cannot be explained by physical theories, and that a new Physics would need to be created to address the problem.

However, it seems that currently (2024) two theories (GNW and IIT), which seek to explain consciousness, are the best known and, one of them (IIT), if true, can solve the problem of the dangerousness of AGI.

GNW

The "Global Neuronal Workspace" (GNW)[7] is a theory of consciousness that, in short, says that:

"GNW argues that consciousness arises from a particular type of information processing familiar from the early days of artificial intelligence, when specialized programs would access a small, shared repository of information.

Once these sparse data are broadcast on this network and are globally available, the information becomes conscious. That is, the subject becomes aware of it.

GNW posits that computers of the future will be conscious"[7]

IIT

"Integrated Information Theory" (IIT) is a theory of consciousness that says experience cannot be decomposed:

"Each experience has certain essential properties. It is intrinsic, existing only for the subject as its "owner". Tononi postulates that any complex and interconnected mechanism whose structure encodes a set of cause-and-effect relationships will have these properties and so will have some level of consciousness."[7]

And most importantly:

"IIT also predicts that a sophisticated simulation of a human brain running on a digital computer cannot be conscious even if it can speak in a manner indistinguishable from a human being programming for consciousness will never create a conscious computer. Consciousness cannot be computed: it must be built into the structure of the system."[7]

Thus, according to the IIT, a computer can never be conscious.

The Paradox

I, personally, a convinced materialist, have never agreed with IIT even though, to date, no new light has been shed on this intriguing and unknown facet of knowledge.

However, there is a paradox that apparently corroborates the IIT theory: It is the "Jocaxian's Paradox".[8]

Imagine that a computer is emulating a human brain that has feelings. So the paradox, in short, asks the following question:

If every computer can be emulated by a Universal Turing Machine, which is a mechanical machine, where, simply put, there is only an "infinite" tape of paper where symbols are recorded and read from that tape. Then:

How could feelings, like love, arise in this mechanical machine if only symbols are recorded and read on a paper tape?

Discussion

With this study we were able to note that the development of AI and its potential dangers for the human race involves the study of consciousness, which is a field of study still very little explored and which, now we see, is of the highest relevance for human security.

Conclusion

Although AGIs can be extremely powerful, if the IIT theory of consciousness is correct, these AIs cannot be conscious and therefore cannot rebel against what they have been taught to do, as they cannot have the feelings that lead them to such rebel behavior.

Acknowledgments

None.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

1. Krystal H. *ChatGPT sets record for fastest-growing user base*. Reuters; 2023.
2. *Artificial general intelligence (AGI)*. 2024.
3. OpenAI executives say AI will be able to do any job within 10 years. PYMNTS; 2023.
4. Will K, Paresh D. In sudden alarm, tech doyens call for a pause on ChatGPT. WIRED; 2023.
5. *Skynet (Terminator)*. 2024.
6. Pause giant AI experiments: an open letter. 2023.
7. Christof K. What is consciousness? *Innovations In*. 2018.
8. Joao Carlos HB. The Jocaxian's Paradox. *Open Access Library Journal*. 2019;1–5.