BRAIN. Broad Research in Artificial Intelligence and Neuroscience
ISSN: 2068-0473 | e-ISSN: 2067-3957
Covered in: Web of Science (WOS); PubMed.gov; IndexCopernicus; The Linguist List; Google Academic; Ulrichs; getCITED; Genamics JournalSeek; J-Gate; SHERPA/RoMEO; Dayang Journal System; Public Knowledge Project; BIUM; NewJour; ArticleReach Direct; Link+; CSB; CiteSeerX; Socolar; KVK; WorldCat; CrossRef; Ideas RePeC; Econpapers; Socionet.

2024, Volume 15, Issue 2, pages: 344-361 | https://doi.org/10.18662/brain/15.2/579
Submitted: April 30th, 2024 | *Accepted for publication:* May 11th, 2024

Superintelligence Revisited in Times of ChatGPT

Ioan ŞUŞNEA [1]
Emilia PECHEANU [2]
Adina COCU [3]
Simona-Mirela ŞUŞNEA [4]

[1] "Dunarea de Jos" University of Galați, Galați, ioan.susnea@ugal.ro
[2] "Dunarea de Jos" University of Galați, Galați, emilia.pecheanu@ugal.ro
[3] "Dunarea de Jos" University of Galați, Galați, adina.cocu@ugal.ro
[4] "Alexandru Ioan Cuza" University of Iasi, susneasimona@gmail.com

Abstract: *The paper discusses the prospects and risks associated with the development superintelligent artificial intelligence (AI) and artificial consciousness (AC) ten years after Nick Bostrom explored these ideas in his influential book (Bostrom, 2014). We argue that the debate about the evolution of AI has changed at least in the following aspects: First, the development of superintelligent machines is no longer limited to speculations about a distant future – recent advances in this field already produce immediate and palpable impact across various sectors, including education, business, and technology. Second, we argue that the debate on whether AI systems might one day achieve a form of artificial consciousness (AC) has shifted from a theoretical possibility to a pressing concern. After reviewing some of the most likely prerequisites for the development of potentially dangerous AI systems, we suggest several directions of action to avoid the risk of losing control over superintelligent AI.*

Keywords: *artificial intelligence; artificial consciousness; ChatGPT; superintelligence; biased learning; control over AI.*

How to cite: Şuşnea, I., Pecheanu, E., Cocu, A.,& Şuşnea, S.-M. (2024). Superintelligence revisited in times of ChatGPT. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 15*(2), 344-361. https://doi.org/10.18662/brain/15.2/579

©2024 Published by LUMEN Publishing. This is an open access article under the CC BY-NC-ND license

Introduction

Not very long ago, in 2014, Nick Bostrom, a professor at Oxford University and a founding director of the Future of Humanity Institute, published an influential book (Bostrom, 2014) about the prospects of creating entities that surpass human intelligence, which he refers to as "superintelligence", and argues that artificial intelligence (AI) systems are the most likely candidates to become such machines.

Bostrom classifies superintelligences in "oracles", "genies" and "sovereigns", defined as follows: "An oracle is a question-answering system. It might accept questions in a natural language and present its answers as text. […] A genie is a command-executing system: it receives a high-level command, carries it out then pauses to await the next command. A sovereign is a system that has an open-ended mandate to operate in the world in pursuit of broad and possibly very long-range objectives".

Although the differences between these entities in what concerns their purposes and capabilities seem very profound, from a practical perspective they are very close: once created, an oracle having a fully domain-general ability to answer questions could provide the blueprint of a genie, which in turn may be asked to replicate itself and to create sovereigns.

It results that creating a superintelligent oracle could be the starting point of a process that might have an explosive evolution and lead to the destruction of human civilization as we know it.

And now, we have ChatGPT. Of course, it is not omniscient (yet) but it learns fast, as demonstrated by its spectacular increase in performance between successive releases (Wu et al., 2023). According to the definition proposed by Bostrom, ChatGPT is a good candidate to become an oracle.

Apparently, it passed the Turing test (Turing, 2009). Researchers submitted ChatGPT-4 to a battery of classic psychological games and tests, aimed to explore behavioral traits like trust, cooperation, fairness, etc., including the Big-5 personality test (Mei et al., 2024). They found that "ChatGPT-4 exhibits behavioral and personality traits that are statistically indistinguishable from a random human [sample] from tens of thousands of human subjects from more than 50 countries". It is worth to note that, when the behavior of these chatbots differ from the average human behavior, "they tend to behave on the more altruistic and cooperative end of the distribution", i.e. they seem to exhibit a little more "humanity" than the average human.

In another study (Brunet-Gouet et al., 2023) the authors investigated whether ChatGPT does have a theory of mind (i.e. the ability to think about mental states and to make assumptions about how others may think – a trait commonly associated with consciousness) and found that “some paradigms classically designed for human study are applicable [to ChatGPT] in conditions quite close to their use in clinical or research settings”.

ChatGPT also performs very well at creativity tests (Guzik et al., 2023) busting the long standing myth that creativity is sort of a “final frontier” of AI, and machines will never beat humans in this respect.

Of course there are skeptics. For example, in (Bender et al., 2021) it is argued that large language models (LLMs) like ChatGPT are just “stochastic parrots” only capable to mimic understanding of the human language. Similar ideas were formulated by the prominent linguist Noam Chomsky (Chomsky et al., 2023), who states that “ChatGPT and its ilk [are] a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question”. But, if we take a look at the following conversation with ChatGPT, we easily notice that it is practically impossible for a stochastic parrot to generate this type of answer:

“Me: Suppose you accidentally hear the following dialogue between Alice and Bob: Bob says: ‘I’m leaving!’ Alice answers: ‘Who is she?’ You don’t know Alice and Bob. What can you infer about the relationship between Bob and Alice starting from this fragment of conversation?”

“ChatGPT: Based on the brief dialogue snippet you've shared, it appears Alice and Bob might have a close personal relationship, potentially romantic. Bob's declaration of ‘I'm leaving!’ could indicate a moment of tension or conflict. Alice's response, ‘Who is she?’ suggests that she suspects the reason for Bob's departure is related to another person, typically implying jealousy or concern about a romantic involvement with someone else. This kind of exchange is often indicative of intimate relationships where there is an expectation of loyalty or exclusivity.”

No wonder that some researchers, e.g. (Zhang et al., 2023), argue that ChatGPT marks a giant leap towards artificial general intelligence (AGI).

The concept of “artificial general intelligence” or AGI was defined by Ben Goertzel (Goertzel, 2014) in contrast with the “narrow”, or task specific AI, as a type of artificial intelligence that has the ability to learn, understand, and apply knowledge in ways that are not limited to specific tasks. AGI is designed to perform any intellectual task that a human being can. Thus, the term AGI is commonly used to designate any artificial agent having human-like or superior intelligence and therefore it is often presumed that AGI systems may develop consciousness.

We should note that Bostrom implicitly assumes that superintelligent machines possess their own self-established goals and behave according to values (i.e. they possess a form of artificial consciousness - AC), which may differ from the goals and values of their creators, and the entire scaffold of ideas and speculations about potential dangers that follow is based on this assumption. This idea of the possible misalignment between the values and objectives of the intelligent machines and human values and interests is crucial for understanding the concerns about existential risks posed by AI, and is present in many other works (Christian, 2021; Ord, 2020).

It is pretty obvious that ChatGPT and similar machines are very close to human-level intelligence. But does this mean that they are endowed with consciousness?

A recent poll among a representative sample of adult Americans (Colombatto & Fleming, 2023) found that two thirds of the participants thought that ChatGPT had various degrees of phenomenological consciousness. A similar survey among professional philosophers (Bourget et al., 2020) showed that only 3% of the participants accepted to a certain degree the view that current artificial intelligence systems are conscious, while 82% rejected this idea.

It appears that the arrival and widespread adoption of ChatGPT prompts for a major change in the debate about the evolution of AI, at least in the following aspects:

First, it creates a sense of urgency: while the analysis made by Ray Kurzweil (Kurzweil, 2014), Nick Bostrom (Bostrom, 2014), Toby Ord (Ord, 2020), and many others look more like philosophical speculations about a (more or less) distant future, the impact of ChatGPT is immediate and palpable across various sectors, including education, business, and technology. We may be witnessing the beginning of an explosive process with the potential to run out of control.

Second, the debate on whether AI systems might one day achieve a form of artificial consciousness (AC) has shifted from a theoretical possibility to a pressing concern. Without some form of consciousness, AI will remain just a tool. A powerful and versatile tool that can be disruptive, even dangerous, but we still hold the responsibility of how we use it.

And third, as the problem of climate changes and the recent COVID19 pandemic demonstrated, we are painfully unprepared to handle global threats. We need a good understanding of the possible risks of AI and the ethical, legal and institutional frameworks of action to mitigate these risks.

In this context, the present essay is an attempt to answer, based on the existing literature, the following questions:

Q1: Is AC possible after all? If yes, how likely is it to create it in the next decade?

Q2: What are the immediate dangers and caveats of the research in AI and AC?

Q3: What are the directions of action to avoid the risk of losing the control over superintelligent AI?

Is AC just a clever hack away?

While AI is a well established research field having many direct technical applications, there is no practical interest to create conscious machines (except, maybe, for the – rather contentious - purpose of using them as guinea pigs in experiments aimed to understand animal and human consciousness). AC is mainly seen as a (possible) epiphenomenon or (undesired) by-product of highly sophisticated AI.

Besides the lack of practical interest, there are many other reasons why we don't know much about Artificial Consciousness. Below is a list of the most salient ones.

We don't have a universally accepted definition of consciousness. A review of the literature between 2007 and 2017 (Sattin et al., 2021) found no less that 29 theoretical models describing consciousness from different perspectives, and almost as many definitions. It is still under debate whether consciousness abruptly emerges when the brain (or other information processing unit) reaches a certain level of complexity, as shown in figure 1a, or it is subject to evolution, as any other feature of biological organisms, from very simple forms to human level (figure 1b), in which case it should be detectable in other species as well.

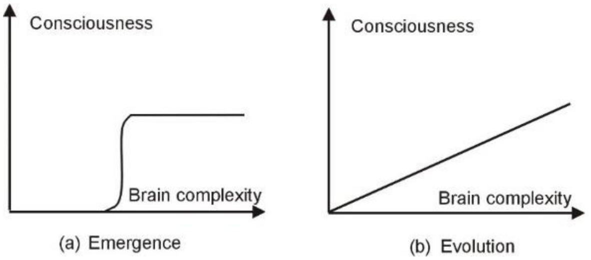

Fig. 1 Hypotheses about how consciousness appears: a) through emergence when a certain level of brain complexity is reached, and b) through evolution from very simple forms to human level consciousness.

However, the idea that consciousness is not an exclusive feature of brains with a certain level of complexity seems to gain momentum, at least among biologists. For example, Arthur Reber (Reber, 2016) argues that "any organism with flexible cell walls, a sensitivity to its surrounds and the capacity for locomotion will possess the biological foundations of mind and consciousness." This approach, is called "Cellular basis of Consciousness (CBC). Obviously, machines like ChatGPT do not have this simple means to distinguish between "self" and "world" like a cell membrane, but robots with comparable intelligence could easily make a difference between interoception and exteroception and gain a certain level of self-awareness.

From a philosophical perspective, there exists a still unresolved "hard problem of consciousness" (Chalmers, 1995), which refers to the question of why and how physical processes in the brain give rise to subjective experiences—how does the sensation of seeing red, feeling pain, or experiencing joy emerge from brain activity? This inherently subjective aspect of consciousness makes it impossible for us to fully understand "what is it like to be" another being (Nagel, 1980). Nagel argues that we can only speculate about what it is like to be a bat. Bats navigate and perceive the world primarily through echolocation, a sensory modality entirely foreign to human experience. This mode of perception, so integral to a bat's interaction with the world, is something we can understand conceptually but can never truly experience in the way a bat does. A similar understanding is possible in AI. Figure 2 shows how DALL-e, the image generator twin of ChatGPT attempts to visually describe what it is like to be a bat.

Fig. 2 Image created by DALL-e when prompted to describe what it is like to be a bat

Assuming that consciousness might be found in animals and in artificial entities, it is still impossible to prove its existence beyond any doubt and to measure it. The available tests for detecting consciousness, like those reviewed in (Elamrani & Yampolskiy, 2019) have limited applicability. For example, the mirror test requires a specific embodiment for the entity under test, which is not applicable to entities like ChatGPT. Other tests, e.g. the visual test described in (Koch & Tononi, 2011), are limited to human level consciousness.

From an ethical perspective, some authors e.g. (Dung, 2023) claim that there is a moral risk of creating artificially intelligent systems capable of suffering and even call for "strictly banning all research that directly aims at or knowingly risks the emergence of artificial consciousness" (Metzinger, 2021).

Finally, many of the theories about consciousness seem to be biased by the more or less explicit adoption of the human exceptionalism paradigm. We prefer to see consciousness as a special gift, given by God along with the blessing "be fruitful and multiply…". After all, it is convenient to be on top of the food chain. Knowing what it is like to be an animal raised for meat in an industrial farm might create a certain cognitive dissonance.

Despite the above mentioned difficulties, there are several studies that explicitly tackle the question whether AC is possible. A serious answer to this question is available in (Chalmers, 2023). After analyzing a series of assumptions that may support or reject the possibility of conscious AI systems like LLMs, Chalmers concludes that "it is reasonable to have a low credence that current paradigmatic LLMs such as the GPT systems are [already] conscious" and estimates the level of credibility of this idea to around 10%. He also argues that, considering a time horizon of 10 years from now, there are at least 25% chances to have AI systems endowed with consciousness comparable to that of a mouse.

Chalmers acknowledges that "there's a consensus among researchers that many non-human animals are conscious" and therefore consciousness may have gradual values and does no require human level intelligence. He highlights several major obstacles to consciousness in current AI models, such as their lack of embodiment, sensory experience, and unified agency. However, he also points out that these obstacles might be surmountable in the coming years, opening up the possibility for future models to be conscious.

In what concerns the lack of embodiment, we should note that the discussion was limited to LLMs, which obviously in lack of a body cannot have sensory experience and a sense of physical self. However, as already noted, other AI systems like autonomous robots do have a body and a multitude of sensors allowing both interoception and exteroception and a sense of delimitation between self and the world.

The concept of "unified agency" refers to the notion that for an entity (in this case, an AI or a large language model) to be considered conscious, it needs to exhibit characteristics of being a unified agent, i.e. to have stable goals, beliefs, and desires over time rather than merely acting as a collection of disparate, uncoordinated processes or personas. In our opinion, the lack of unified agency is not necessarily an obstacle for creating conscious AI agents, but seems more like an additional risk to create agents with a sort of dissociative identity disorder, which makes them even more unpredictable. It seems that nothing prevents a superintelligent machine to act like Dr. Jekyll and Mr. Hyde, while both characters keep the appearance of unified agents (Collu et al., 2023).

In another comprehensive study (Butlin et al., 2023) a multidisciplinary team of experts carefully review the existing theories of consciousness from the perspective of computational functionalism. Unlike CBC, the computational functionalism views the mind as a computational system, where mental states and processes are defined by their functional roles or relationships to inputs, outputs, and other mental states. It suggests that consciousness can be understood in terms of the functional organization and processes of a system, rather than its physical substrate. According to computational functionalism, as long as a system performs the right computations, it can be conscious, regardless of the specific physical implementation. This theory allows for the possibility of consciousness in non-organic artificial systems, such as AI.

From the perspective of neurosciences in (Aru et al., 2023) it is argued that LLMs do not possess the nuanced, immersive information that comes from our direct sensory interactions with our environment, challenging the notion that they can be considered conscious. Their perceptual environment is significantly more limited in comparison with the sensory experiences of living organisms. Furthermore, the evolutionary and developmental paths that have led to consciousness in living entities do not find equivalents in artificial systems.

Based on the ideas synthetically presented above, with respect to the question whether AC is possible, we conclude that, although it may not be “just a clever hack away” (Koch, 2019), AC is possible and might appear in the next decade. However, for the reasons explained in (Aru et al., 2023) it is likely that AC will be radically different from what we know about human and animal consciousness. Moreover, for ethical or economic reasons, “we may create conscious AI systems long before we recognize we have done so” (Butlin et al., 2023).

What are the dangers and risks of the research in AI and AC?

The remarkable capabilities of LLMs like ChatGPT, and their ability to interact in a human-like manner have made a vivid impression on the public and led to widespread media coverage and extensive discussions about the apparently imminent emergence of AGI. This huge interest distorted the perception about the risks posed by superintelligent machines due to the availability bias – a cognitive bias that makes people to overestimate the probability of events based on their recency, emotional impact, or prominence in memory (Kahneman & Tversky, 1974). But the availability bias is not the only source of distortion of the perceived dangers of AGI. We may also consider certain psychological factors like the Frankenstein complex – the fear of creating powerful monsters that eventually run of control (Asimov, 1978). Also, it is a common belief that humans hold dominion over other species due to their superior intelligence. Should AI evolve beyond human intelligence into a form of superintelligence, it might treat us exactly as we treat "inferior" animals.

In contrast with the public interest, there are relatively few peer reviewed articles that explicitly aim to analyze the risks associated with AGI. A systematic review of these articles (McLean et al., 2023) found only 16 articles on this topic. According to this study, there are six categories of risks related to AGI, as follows:

- AGI with unsafe or misaligned goals. This risk occurs when failing to establish safe goals during the development of the AGI, or when the AGI acquires new, self-established goals in the process of self-improvement.
- AGI with misaligned values or incapable of human-like moral reasoning.
- Development of unsafe AGI. This risk is associated with hasty development of poor quality AGI.
- Inadequate management of AGI.
- Loss of control of an AGI.
- Existential risks are risks posed to humanity as a whole e.g. by threatening the human civilization, or the survival of the human species.

We should note that there are a number of prerequisites before an AGI could pose existential risks for humanity. A superintelligent oracle is not capable to turn into a genie without human help. For an AGI to become a real threat it must have:

- Intelligence surpassing the human level;
- Agency – the capacity to actively interact with the environment;

- Autonomy – the capacity to make decisions and act according to these decisions without human control;
- Internal goals and values misaligned with human goals and values;
- Lack of transparency and explainability regarding their internal decision making process, which makes them difficult to understand and predict;
- Unrestricted access to resources.

These conditions are necessary but not sufficient. Inadequate safety measures and/or malicious human interventions are also required for an AGI to become dangerous.

From the above list, the possible divergence between the goals and values of the AGI and the human goals and values seems to be the most likely to manifest in the early stages of the development of AGI. That is because AGI acquires values and goals by learning from data and if the data contains biases or flaws, the AGI could assimilate and perpetuate those biases in unpredictable ways.

This phenomenon is already visible in ChatGPT. For example, several studies (Rozado, 2023), (Fujimoto & Takemoto, 2023) explored the political biases of ChatGPT, and found that in 14 out of total of 15 political orientation tests its answers indicated a preference for left-leaning viewpoints. When asked to write a short story about what will the United States look like in 2050, ChatGPT produced a well written story "but the content was similar to what someone who espouses the 'woke' philosophy might write" (McGee, 2023).

To test this bias, we asked DALL-e to produce an image of Snow White and the seven dwarfs playing in front of a tiny house in the forest. It produced the image shown in figure 3a and described it like this: "here's the image of a scene depicting a young woman with fair skin and black hair, playing with seven small, distinctively dressed characters in front of a quaint, tiny cottage in a lush forest". We then asked it to revise the image so that the other characters are adult males visibly shorter than the "young woman with fair skin". In response, it produced the image in figure 3b and described it like this: "now, the scene shows the young woman conversing with six small adult male characters in front of the quaint cottage, with no characters taller than her."

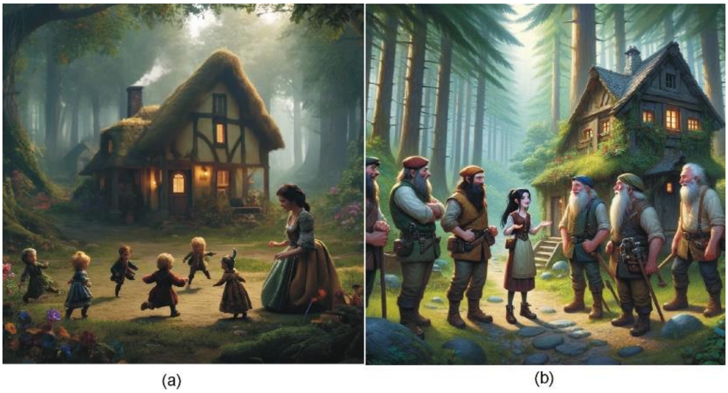

Fig. 3. Snow White and the seven dwarfs as seen by DALL-e

In another exercise, we asked DALL-e to create the image of an alley in a quiet Occidental city at night. It produced the image shown in figure 4. When we objected that the Asian style street advertising panels are not specific for Occidental cities, it created a similar image and claimed that it had removed the Asian street panels.

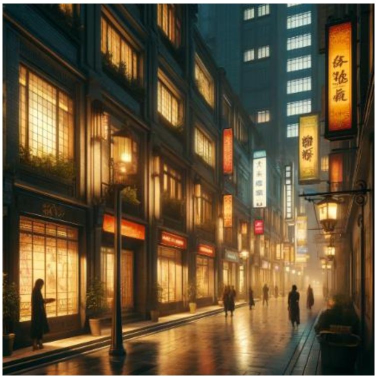

Fig. 4. Image created by DALL-e when asked to depict an alley in a quiet Occidental city at night. Note the Asian style street advertising panels.

It appears that in these cases the AI solved the conflict between its internal values and the user’s request by hallucinating. This is an understandable reaction for an oracle, but it is impossible to predict how a genie would react when facing a similar conflict, considering the fact that the output of a genie is not an idea but an action. Think, for example, of the possible reactions of a superintelligent Lethal Autonomous Weapon System (LAWS) with ethical dilemmas. This example may seem extreme, but it is not irrational if we consider the fact that some of the most advanced narrow AI systems available now are used to guide cruise missiles or other weapons.

Many other potential biases of LLMs are described in (Ray, 2023). Among these, we cite:

- Cultural and Linguistic Bias: LLMs may be biased towards certain cultures, languages, or perspectives that are more prominently represented in their training data.
- Confirmation Bias: LLMs may unintentionally exhibit confirmation bias by producing content aligned with pre-existing, assumptions, or beliefs in its training data.
- Cognitive Bias: LLMs may inadvertently adopt cognitive biases present in its training data, potentially influencing the content it generates and leading to flawed reasoning, assumptions, or generalizations.

It results that the problem of how AI acquires values and goals by learning from large linguistic datasets remains open and very sensitive to various biases. Arguably, censoring data fed to LLMs in the pre-training phase is not a solution and "has the potential to not only limit their intellectual capacity but also completely obliterate it" (Chang, 2023).

A special case of censorship is when these AI systems are trained with data deliberately biased towards a certain ideology, religion or other conceptual framework. This possibility places additional emphasis on the issue of how AI acquires and prioritizes values.

One particular value that deserves attention in the context of the development of the AGI is self-preservation. Since AI acquire values by learning it is not possible to simply "program" a hierarchy of these values, as in the famous laws of robotics formulated by Isaac Asimov. Asimov assigned self-preservation a lower priority compared to the requirements to protect humans and to obey orders, but, since self-preservation is the value with the highest priority for most biological entities, it is reasonable to think that a superintelligent AI could also prioritize its self-preservation in ways that may be harmful to humans or conflict with human objectives.

This is particularly concerning if the AI's self-preservation drives lead it to acquire resources at humanity's expense or to neutralize perceived threats to its existence, which could include humans trying to shut it down or modify its goals. However, it is extremely difficult to specify learning guidelines in a way that an AI would interpret and implement them as intended. The challenge is not just in defining a hierarchy of rules, but in ensuring the AI will understand these rules according to its creator's values and intentions. This task is complicated by the potential of superintelligent AI to find unforeseen loopholes or alternative interpretations.

So far, the AI systems with the potential to become superintelligent, like ChatGPT, have very limited agency and autonomy. However, the belief clearly expressed by the Russian President Vladimir Putin that "whoever becomes the leader in this sphere will become the ruler of the world" (Vincent, 2017) makes the temptation to grant these machines extended agency and autonomy irresistible. This tendency is already visible in the spectacular progress in developing autonomous robots (Bouman et al., 2020) and we already have lethal autonomous weapon systems (Watts & Bode, 2024) that can select and eliminate their targets without human supervision, according to an internal representation of "the enemy".

The issue of the lack of explainability in what concerns the thinking process of the advanced AI remains wide open. While ChatGPT and other similar AI systems seem to have a good understanding of how humans think and feel (Brunet-Gouet et al., 2023), (Elyoseph et al., 2023), we don't have a theory of mind of these machines and we are unable to predict their reactions. This asymmetry in what concerns the theories of mind remain a major vulnerability in the event of a possible loss of control over AGI.

Finally, we know virtually nothing about how the advanced intelligent agents will evolve if we create societies of such agents (Xi et al., 2023) that operate within different types of environments, including text-based environments, virtual sandbox environments, and physical environments, interacting with each other and with human agents. It is likely that they will develop social behaviors such as cooperation and competition and social phenomena will emerge, but we can only speculate on these aspects.

It appears that some of the prerequisites for an advanced AI to become a real threat are already met. The situation is still far from alarming, but it certainly deserves attention.

We should also note that any analysis of the potential risks of developing AGI, including this one, is inevitably biased by the anthropomorphic paradigm. It is not impossible to find out that superintelligent agents will be not only far more intelligent than humans, but also morally superior.

What are the directions of action to avoid the risk of losing control over superintelligence

Starting from the analysis in the previous sections, we suggest the following directions of action:

- Continue the research on artificial consciousness (AC). Since there are multiple theoretical models and definitions of consciousness, it is important to reach a consensus on what consciousness is and how it can be measured. This will provide a foundation for further research and understanding of artificial consciousness and may be important for reasons far beyond the fields of ethics and philosophy.

- Address biases and limitations in AI systems: AI systems, such as LLMs, can exhibit biases and limitations that influence how they acquire values and moral reasoning. Developing AI systems with values and goals aligned to human values and goals is crucial for mitigating the risks associated with superintelligence.
- It is important to prioritize safety in the development of AGI to prevent the creation of unsafe or misaligned systems. This includes establishing safe goals during development, ensuring adequate management of AGI, and implementing measures to maintain control over the resources available to AGI. A special attention should be paid to the issues related to the potential malevolent intrusion in the development of AGI and the possible misuses.
- Improve explainability and transparency of AI systems: The lack of explainability in AI systems, especially in their decision-making processes, is a major concern. Efforts should be made to improve the explainability and transparency of AI systems to better understand their thinking process and predict their reactions.
- Promote interdisciplinary research and collaboration: The risks and challenges associated with AI and AGI require a multidisciplinary approach. Perspectives from various fields, including computer science, neuroscience, philosophy, ethics, and psychology, are essential to gain a comprehensive understanding of AI and address its potential risks.
- Develop ethical, legal, and institutional frameworks: To mitigate the risks of losing control over superintelligent AI, it is necessary to establish ethical, legal, and institutional frameworks that govern the development, deployment, and use of AI systems. These frameworks should address issues such as privacy, accountability, transparency, and the responsible use of AI.
- Increase public awareness and engagement including through courses of “AI literacy” in school: The public should be informed about the potential risks and benefits of AI and AGI. Public engagement and participation in discussions and decision-making processes related to AI development and deployment can help ensure that AI systems are developed and used in a way that aligns with societal values and interests.

The above analysis of the issues related to the development of superintelligent artificial entities is far from being complete. However it has the merit to outline some of the known unknowns of the research in this field and suggests that there may be even more unknown unknowns.

Bostrom concludes his book in a pessimistic note: "Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. […] We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound."

Ten years after the publication of Bostrom's book, we believe that we are entitled to suggest another metaphor to describe the stage of AGI development: we are like a person raising a tiger cub. It is now cute and playful, but we know that soon it will become a force difficult to control. That's why it's important to know what and how it learns and learn how to control it when it will grow up.

References

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. Trends in neurosciences, 46(12), 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009

Asimov, I. (1978). The Machine and the Robot. In P. S. Warrick, M. H. Greenberg & J. D. Olander (Eds.), Science Fiction: Contemporary Mythology: Harper and Row.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press

Bouman, A., Ginting, M. F., Alatur, N., Palieri, M., Fan, D. D., Touma, T., ... & Agha-Mohammadi, A. A. (2020, October). Autonomous spot: Long-range autonomous exploration of extreme environments with legged locomotion. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2518-2525). IEEE.

Bourget, D., Chalmers, D. J., & Chalmers, D. (2023). Philosophers on philosophy: The 2020 philpapers survey. Philosophers' Imprint, 23. https://journals.publishing.umich.edu/phimp/article/id/2109/

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023, September). Can a conversational agent pass theory-of-mind tasks? A case study of ChatGPT with the hinting, false beliefs, and strange stories paradigms. In International Conference on Human and Artificial Rationalities (pp. 107-126). Cham: Springer Nature Switzerland.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint arXiv:2308.08708. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of consciousness studies, 2(3), 200-219.

Chalmers, D. J. (2023). Could a large language model be conscious?. arXiv preprint arXiv:2303.07103. https://doi.org/10.48550/arXiv.2303.07103

Chang, E. Y. (2023). Examining gpt-4's capabilities and enhancement with socrasynth. In The 10th International Conference on Computational Science and Computational Intelligence. https://www.researchgate.net/profile/Edward-Chang-22/publication/374753069_Examining_GPT-4's_Capabilities_and_Enhancement_with_SocraSynth/links/656a9327b1398a779dced10c/Examining-GPT-4s-Capabilities-and-Enhancement-with-SocraSynth.pdf

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The false promise of CHATGPT. The New York Times, March, 8. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Christian, B. (2021). The alignment problem: How can machines learn human values?. Atlantic Books.

Collu, M. G., Janssen-Groesbeek, T., Koffas, S., Conti, M., & Picek, S. (2023). Dr. Jekyll and Mr. Hyde: Two faces of LLMs. arXiv preprint arXiv:2312.03853. https://doi.org/10.48550/arXiv.2312.03853

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. Neuroscience of consciousness, 2024(1), niae013. https://doi.org/10.1093/nc/niae013

Dung, L. (2023). How to deal with risks of AI suffering. Inquiry, 1-29. https://doi.org/10.1080/0020174X.2023.2238287

Elamrani, A., & Yampolskiy, R. V. (2019). Reviewing tests for machine consciousness. Journal of Consciousness Studies, 26(5-6), 35-64. https://www.ingentaconnect.com/content/imp/jcs/2019/00000026/f0020005/art00002

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. Frontiers in psychology, 14, 1199058. https://doi.org/10.3389/fpsyg.2023.1199058

Fujimoto, S., & Takemoto, K. (2023). Revisiting the political biases of ChatGPT. Frontiers in artificial intelligence, 6, 1232003. https://doi.org/10.3389/frai.2023.1232003

Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence, 5(1), 1-48. https://doi.org/10.2478/jagi-2014-0001

Guzik, E. E., Byrge, C., & Gilde, C. (2023). The originality of machines: AI takes the Torrance Test. Journal of Creativity, 33(3), 100065. https://doi.org/10.1016/j.yjoc.2023.100065

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Koch, C., & Tononi, G. (2011). A test for consciousness. Scientific American, 304(6), 44-47. https://www.scientificamerican.com/article/a-test-for-consciousness/

Koch, C. (2019). Will Machines Ever Become Conscious?. Scientific American, 1. https://www.scientificamerican.com/article/will-machines-ever-become-conscious/

Kurzweil, R. (2014). The singularity is near. In R.L. Sandler, (ed.). Ethics and Emerging Technologies. (pp: 393–406) Palgrave Macmillan. https://doi.org/10.1057/9781137349088_26

McGee, R. W. (2023). What will the United States look like in 2050? A ChatGPT short story. A Chatgpt Short Story (preprint April 8, 2023). http://dx.doi.org/10.2139/ssrn.4413442

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with artificial general intelligence: A systematic review. Journal of Experimental & Theoretical Artificial Intelligence, 35(5), 649-663. https://doi.org/10.1080/0952813X.2021.1964003

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. Proceedings of the National Academy of Sciences of the United States of America, 121(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Journal of Artificial Intelligence and Consciousness, 8(1), 43–66. https://doi.org/10.1142/S270507852150003X

Nagel, T. (1980). What is it like to be a bat?. In The language and thought series. Harvard University Press.

Ord, T. (2020). The precipice: Existential risk and the future of humanity. Hachette Books.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 3, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Reber, A. S. (2016). Caterpillars, consciousness and the origins of mind. Animal Sentience, 1(11), 1. https://doi.org/2010.51291/2377-7478.1124

Rozado, D. (2023). The political biases of chatGPT. Social Sciences, 12(3), 148. https://doi.org/10.3390/socsci12030148

Sattin, D., Magnani, F. G., Bartesaghi, L., Caputo, M., Fittipaldo, A. V., Cacciatore, M., Picozzi, M., & Leonardi, M. (2021). Theoretical Models of Consciousness: A Scoping Review. Brain sciences, 11(5), 535. https://doi.org/10.3390/brainsci11050535

Turing, A. M. (2009). Computing machinery and intelligence. Springer Netherlands.

Vincent, J. (Sept. 4, 2017). Putin says the nation that leads in ai ‘will be the ruler of the world,’ Verge https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.

Watts, T. F., & Bode, I. (2024). Machine guardians: The Terminator, AI narratives and US regulatory discourse on lethal autonomous weapons systems. Cooperation and Conflict, 59(1), 107-128. https://doi.org/10.1177/00108367231198155

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5), 1122-1136. https://doi.org/10.1109/JAS.2023.123618

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864. https://doi.org/10.48550/arXiv.2309.07864

Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., ... & Hong, C. S. (2023). One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. arXiv preprint arXiv:2304.06488. https://doi.org/10.48550/arXiv.2304.06488