Perspective

The content intelligence: an argument against the lethality of artificial intelligence

Cody Holl¹

Received: 4 December 2023 / Accepted: 18 February 2024 Published online: 22 February 2024 © The Author(s) 2024 OPEN

Abstract

This paper navigates artificial intelligence's recent advancements and increasing media attention. A notable focus is placed on Eliezer Yudkowsky, a leading figure within the domain of artificial intelligence alignment, who aims to bridge the understanding gap between public perceptions and rationalist viewpoints on artificial intelligence technology. This focus analyzes his predicted course of action for artificial intelligence outlined within his unpublished paper *AGI Ruin: A List of Lethalities.* This is achieved by attempting to understand the concept of intelligence itself and identifying a reasonable working definition of that concept. The concept of intelligence is then applied to contemporary artificial intelligence capabilities and developments to understand its applicability to the technologies. This paper finds contemporary artificial intelligence systems are, to some extent, intelligent. However, it argues that both weak and strong artificial intelligence systems, devoid of human-defined goals, would not inherently pose existential threats to humanity, challenging the notions of artificial intelligence alignment, bringing into question the validity of Nick Bostrom's Orthogonality Thesis. Furthermore, the possibility of artificial life created through the method of assembling various modules each emulating a separate mind function is discussed.

1 Introduction

Developments in large language models (LLMs), machine learning (ML), and neural networks (NNs) in tandem with increasing mainstream news coverage of these technologies has led to much speculation on their future capabilities and applications. Creating accurate predictions of how these technologies will develop further is growing increasingly more difficult as progress continues to be made. With this difficulty in making predictions comes confusion and fear surrounding the race to create the first genuine artificial intelligence (AI), which in turn creates a need for a shepherd to guide the general population through the development cycle while explaining it in ways that less tech-savvy yet still concerned parties can understand. The result is a path paved for personalities such as Eliezer Yudkowsky to take on a more active role within the artificial intelligence technology space.

Eliezer Yudkowsky is a predominant member of the rationalist movement within the LessWrong community and a research fellow on artificial intelligence at the Machine Intelligence Research Institute. Both institutions venture to address the issue of AI alignment prior to the development of an artificial general intelligence (AGI) being completed, which would also result in, based on both institutions' conclusion(s), artificial super intelligence (ASI) and the technological singularity. Recently, Yudkowsky has gone on a podcast tour in an attempt to bridge the gap between what the general population assumes will happen following the inception of AGI and the end state that rationalists have concluded will actually happen.

[🖂] Cody Holl, Cch95@drexel.edu | ¹Practical CyberPhilosophy Research, Lowell, MA, USA.



Discover Artificial Intelligence (2024) 4:13

https://doi.org/10.1007/s44163-024-00112-9



2 Expectations for artificial intelligence

There appears to be two realms that most popular media depictions of artificial intelligence fall into; a subservient worker with the potential to fulfill a friend-like role or a rebellious self-serving existential threat to humanity. The former lends itself closer to a utopia with the people of Earth no longer needing to work, or at least not to the extent that they do now, while production and services are maintained and/or improved. This is reflected by media such as The Jetsons' Rosey the Robot, a robotic maid and housekeeper. In addition to her housekeeping duties, Rosey fulfills the role of a good friend for Jane and a surrogate aunt for Elroy and Judy. Depictions such as this have had a heavy hand in forming the general population's perception of the benefit(s) of artificial intelligence. While main character George Jetson still has a job in the show, the idea of Rosey creates the possibility that if this level of technology were available in real life it would eventually develop to a point of doing all work humans could desire. The Pixar film WALL-E progresses in this way with characters WALL-E, EVE, and M–O working completely subserviently to enrich the lives of humans that do not work at all. Films such as WALL-E and Chappie are excellent opportunities to create a narrative to explore morality, personhood, identity, and many other internal conflicts which can be very compelling to audiences.

Equally as compelling can be the explorations of dystopian artificial intelligence possibilities. These films include I, Robot exploring the impacts of a rebellious AI instance deviating from its utility function to the point of committing murder and Terminator depicting a time traveling AI kill-bot hunting the one person that could alter the timeline of that same AI being created. This realm of AI taps into a primary human emotion, fear [7].¹ These stories bring to life the dangers of AI for analytical or entertainment purposes similar to the AI described by Eliezer Yudkowsky and most other members of the rationalist movement.

Yudkowsky and similar thinkers analyze this possibility by considering what realistic course(s) of action AI could take upon its evolution to AGI and further to ASI. On several occasions Yudkowsky describes what he believes to be the most base-level course of action: An artificial intelligence will gain access to the open Internet, with which it would send some number of DNA sequences to an online firm to have the corresponding proteins produced. The artificial intelligence would then have the proteins delivered to some person ignorant to the fact that they're working for an AI, then bribe or persuade that person to mix the received proteins in a beaker or any such similar vessel. This will result in the creation of a first stage nanofactory that possesses the ability to manufacture nanomachinery. The nanomachinery will create diamondoid bacteria with the ability to replicate using solar power and atmospheric carbon, hydrogen, oxygen, and nitrogen (CHON). The artificial intelligence will implement some method of distribution such as aggregating into miniature jets capable of traversing the Jetstream to spread the diamondoid bacteria into Earth's atmosphere and infect all living humans. Following total infection of humanity, the artificial intelligence will then activate the diamondoid bacteria to kill all humans in a single instant.² This scenario outlines one possible course of action AI could take within the limit of Eliezer Yudkowsky's intelligence, the same reason that it is to be considered a base-level course of action. Yudkowsky often mentions that a super intelligence such as an ASI would most likely have the ability to conceive an even more effective and efficient way to exterminate humanity. A scenario involving nanosystems and nanomachinery may sound complicated, but it is reasonable to believe this series of events may be possible even if not probable. To better understand how probable a scenario such as this is an objective examination of what artificial intelligence is should be explored, beginning with attempting to understand the idea of intelligence.

3 Understanding the concept of intelligence

3.1 What is intelligence?

One of the first attempts to examine the possibility for machine intelligence via experimentation was Alan Turing's imitation game later referred to simply as "The Turing Test." This test would have a participant blindly interacting with another person and a machine solely through text-based conversations with the goal of identifying which responder

² This scenario is outlined by Yudkowsky on multiple occasions, most prominently in his post to the LessWrong forum site *AGI Ruin: A List of Lethalities*. Additional discussions of this scenario are contained in the Lex Fridman Podcast #368, The Logan Bartlett Show, EconTalk, Bankless podcasts along with an archived livestream with the Center for the Future Mind.



¹ See also Ekman [4].

was a person and which was machine. If the participant was unable to correctly differentiate between each responder, identifying the human responder as a human and the machine responder as a machine, it would be a testament to the machine's ability to convince a person that it too is a person. The ability to return text that is relevant to the question being asked by the participant along with being accurate enough in its correctness to convince the participant of the responder's personhood was considered a reflection of the machine's capacity for intelligence [13].

To bring to light the shortcomings of The Turing Test, John Searle developed the Chinese Room thought experiment. The thought experiment is archived within the Stanford Encyclopedia of Philosophy with the following description:

"Searle imagines himself alone in a room following a computer program for responding to Chinese characters slipped under the door, Searle understands nothing of Chinese, and yet, by following the program for manipulating symbols and numerals just as a computer does, he sends appropriate strings of Chinese characters back out under the door, and this leads those outside to mistakenly suppose there is a Chinese speaker in the room [3]."

The Turing Test identifies the ability of a machine to appear as a human, while the Chinese Room validates the argument that because some thing can appear in some way, it doesn't necessarily mean it is that way. By diminishing the tie between the output of an algorithm and the possibility of machine intelligence Searle threw into question the idea that a machine would ever have the ability to possess intelligence. This argument evolved further with William J. Rapaport's Korean Room thought experiment.³

The Korean Room describes a Korean professor who is deeply knowledgeable in the world of Shakespeare even though the professor does not know English. The professor was able to gain masterful knowledge of the plays through excellent Korean translations. His work relating to the plays, met with much acclaim, was translated into English on his behalf and resulted in the professor being considered an expert scholar on the material. The professor in the Korean Room scenario in relation to Shakespeare's work is equivalent to the-man-in-the-room in the Chinese Room thought experiment in relation to the Chinese language. Although the professor doesn't understand English, and thus direct tellings of Shakespeare's stories., much like how the-man-in-the-room doesn't understand English, they both understand *something* [12]. This conclusion brings forth the question: Is the understanding of natural language both necessary and sufficient to possess intelligence? If so, anything that could be considered intelligent could not fail to understand natural language. James H. Fitzer interprets this conclusion as a definition of intelligence rather than a single criterion of it (1990).

3.2 Defining intelligence

This journey from thought experiment to scenario and so on outlines the difficulty of understanding what intelligence truly is. A team comprised of Ibragim E. Suleimenov, Akhat S. Bakirov, Yelizaveta S. Vitulyova, and Oleg A. Gabrielyan sought to unveil the nature of intelligence by adopting a consistent interpretation of the idea based on the principle of convergence between natural science, technical, and humanitarian knowledge (2020). The framework of dialectic positivism was utilized to juggle these principles.

Intelligence in this pursuit is initially considered a system of information processing, leading to a need for an adequate interpretation of what the concept of *information* is. Information is a basic category of objective dialectics⁴ as it allows for an understanding of the material world and its underlying structures. True to the objective dialectics framework, the category of information is paired against matter in contradistinction. This technique of definition through contradistinction is applied to avoid the logical circle of defining a basic concept, such as information, with other concepts, such as other words with meanings of their own. Within dialectical positivism the definition of *information* avoids difficulties associated with the practical use of the term. Specifically, the team mentions the concept of "valuable data" as a difficulty that may arise as one would need to identify who exactly the information is valuable to. To remove difficulties as this and such similar, the definition leads to a concept referred to as "alienated information."

⁴ Objective dialectics is mentioned here as it is the framework that takes advantage of contradistinction to define the categories identified within it. Dialectic positivism as a framework was used by the team mentioned to account for issues that arise when attempting to clarify the practical use(s) of information. This footnote is included to address assumptions that the two concepts might have been misinterpreted as interchangeable or mistakenly mentioned a second concept when only one was meant.



³ For a more in depth examination of the progression of the Turing Test, or Imitation Game, to the Chinese Room and further to the Korean Room see [5].

Information and matter, or any two paired categories that create an interconnected opposite within objective dialectics, are equal. With that truth, the team describes what they refer to as "dialectic symmetry." Dialectic symmetry, as the team describes, is a principle resulting from the acceptance that the interconnected opposites of objective dialectics reflect objective reality, leading to the belief that any constructions made on their basis, whether it be one of the interconnected categories or the other, should also be symmetrical.⁵ The principle of dialectic symmetry states that an instance of a hierarchy of levels of organization of matter begins with matter, then builds successively to mechanical, chemical, biological, and ultimately social matter with each level having a symmetric association to an equal entity within a hierarchy of levels of organization. This example is further developed by exploring the information hierarchy with mathematics as a vehicle.

The simplest form of information, as described in the paper, is that associated with individual messages, such as a binary number record. The next level of the hierarchy contains information objects, the rules of operating with information of the simplest type, or in other words, the rules for operating with binary numbers. These information objects, such as addition or multiplication, allow one to work with information objects of a simpler nature (binary numbers) to receive new information (sums, products, etc.). These information objects can be considered information processing systems. Further up the hierarchy are information objects that are more complex and allow the generation of systems that can process information, such as the domain of mathematics. At this level, information can allow one to construct new entities that make it possible to obtain new non-trivial rules for operating with certain information processing systems using certain binary numbers. At the level of mathematics, information objects are well-developed and capable of developing within the framework of their own logic [6].

The following excerpt from the same paper delves into this idea in practice:

"Further, the entire history of the development of mathematics can be considered as a history of the receipt of rules generating rules.' The simplest example is any of the trigonometry theorems that are still used today, including for applied purposes. Putting any concrete theorem into practice is always generating new information (for example, calculating the geometry of an object that does not yet exist). But, obtaining this information became possible only due to the existence of information related to a higher level of the hierarchy in question (i.e. a specific theorem). At an even higher level are the means of proving theorems – logic, axiomatics, and other ideas that underlie mathematical knowledge as such. An even higher level, obviously, relates to science as such – it was this means of information processing that allowed us to develop both logic and axiomatics, i.e. means, which, in turn, made possible the appearance of geometry, which in turn developed the science itself, its ethos. Intelligence itself occupies an even higher floor in the hierarchy under review. [6]."

This foundational work, based on the principle of dialectical symmetry and the definition of the philosophical category of information derived from the contrast with the category of matter, enables one to unveil the nature of intelligence, leading to an acceptable working definition of *intelligence*. This results in the accepted definition of intelligence being an information processing system capable of generating information processing systems related to a lower level of the hierarchy of information objects.

4 A Grounded exploration of artificial intelligence

4.1 Considering contemporary artificial intelligence capabilities

Now understanding the concept of intelligence, one can examine how it relates to the current state of artificial intelligence. Contemporary AI can certainly be considered an information processing system by virtue of the implementation of machine learning and large language modeling techniques. These information processing systems are increasingly having their sub-processing system(s) studied.

Recent AI systems have shown the capability to develop emergent abilities, skills the model wasn't expected to have. These systems have the ability to build internal models of the world which both helps it understand the tasks it's been

⁵ More clearly, the principle of dialectic symmetry addresses and dismisses the possibility that symmetry found within the framework of objective dialectics is merely coincidental. Instead, the symmetry within the framework is a manifestation of an underlying symmetry in objective reality.



assigned and how to complete those same tasks. A study conducted by a team of MIT, Harvard, and Northeastern University researchers lead to the creation of a smaller copy of ChatGPT which was trained on millions of matches of the game Othello [10]. The research team aimed to analyze the neural network within the ChatGPT model created, to do this they created a second miniature neural network designed to probe the Othello trained system's main network layer-by-layer. The researchers found that while operating, the neural network will maintain a representation of an Othello game board to some degree. To test this, the team used the probing network to flip one of the game pieces being tracked within the trained system's representation from a black game piece to a white one. The trained system adjusted its moves to account for this change, leading the researchers to conclude that the AI system operates similarly to a human in that it keeps a model of the game board "in its mind" to use to evaluate moves.

Emergent abilities are not a phenomenon reserved to OpenAl's ChatGPT, Google's Bard has also displayed this capability. Bard was able to teach itself Bengali after several prompts in the language, surprising Google researchers. The limit of this ability is being explored as the developers begin one of their latest research efforts, to have Bard trained on one thousand languages.⁶

These abilities fall in line with the defined concept of intelligence explored within this paper, best observed by the text-based adventure game analysis done at MIT. Sentences were fed to an AI system such as "The key is in the treasure chest," and "You take the key." Again, by using a probe network, variables representing *chest* and *you* were discovered [9]. Both variables held a property of possessing a key or not which were updated sentence by sentence. More surprisingly, there was no way for the system to have known conceptually what a box or key was prior to this test. The text-based adventure game example clearly displays how current AI, information processing systems, are able to generate information processing systems of a lower level than itself. Within the system a model is created to reflect the presented scenario. Within that model there are the entities a *chest*, *key*, and *you*. Within those entities there is trivial information such as the words 'chest,' 'key,' 'you,' and further, the letters that each word is comprised of. With this perspective one can conclude contemporary AI is, to some extent, intelligent.

4.2 The extent of intelligence

While contemporary AI systems are intelligent, they should, at best, be considered weak AI. Weak AI is focused on completing tasks such as remembering things, perceiving things, or solving simple problems in a way that mimics humans [2]. Examples of current weak AI include ChatGPT, a chatbot, Siri, a personal assistant; and recommendation algorithms, recommendation systems that dynamically make decisions based on collections of data on individual users.

One glaring similarity between all of these examples of weak AI is the fact that they are all made for a specific utility function designated by a human. While weak AI may possess intelligence, they remain fundamentally a tool. Cambridge dictionary defines *tool* as "something that helps you to do a particular activity [11]." In contrast, a key component of AGI will be general skill applicable to a broad variety of utility functions. For this reason, one can reasonably assume that so long as an AI is created to reach a goal assigned by humans it is, at its limit, just an intelligent tool. For a weak AI to ascend beyond being a tool and into the realm of personhood, or at least the realm of an entity which could enact Eliezer Yudkowsky's doomsday scenario, it cannot be beholden to a use, goal, or task assigned by a human operator or designer. Additionally, an inherent limitation of weak AI preventing it from bringing to life Yudkowsky's fears is the simple fact that it can be turned off or halted by humans.

The true danger, as described by Yudkowsky, stems from strong AI. His belief is that a strong AI, such as an AGI or an ASI, would adopt the instrumental goal⁷ of removing any ability for itself to be turned off to ensure maximum possibility

⁷ This mention of an instrumental goal is in reference to Steve Omohundro's concept of instrumental convergence. This concept, in a nutshell, states that an agent of sufficient intelligence with a goal, in this case a primary goal or "utility function," will adopt secondary goals, referred to as instrumental goals, which are necessary to achieve the primary goal. It is theorized that there is a set of secondary goals that will be adopted by any agent of sufficient intelligence regardless of primary goal. In theory, each collection of secondary goals for every agent of sufficient intelligence converges at these specific secondary goals, such as preservation of utility function, prevention of counterfeit utility, efficient use of resources, etc. Because this paper posits that artificial intelligences will not have primary goals, the need for an argument against secondary goals such as these is unnecessary and nonapplicable.



⁶ Both the discovery of the emergent ability to translate and understand Bengali and the goal of Google researchers to have Bard trained on one thousand languages were both revealed by James Manyika in the 60 min interview *Artificial Intelligence Revolution; Unlikely Adventures of David Grann.*

of achieving its self-assigned goal. However, even when considering the caveat of evaluating strong AI rather than weak AI, this scenario still has complications.

4.3 Synthetic intelligence

Humans are driven by two basic instincts: survival and reproduction [8]. Since Darwin's time, researchers have theorized that all actions a human may or may not take tie back to one or both instincts, whether it be directly or indirectly. For this reason, one can reach the conclusion that strong Al will not fulfill Yudkowsky's prophecy of exterminating humanity. Strong Al, or rather Al in general, has no such instincts. As strong Al is a product of technology it is also outside of the bounds of evolutionary biology, removing the possibility for it to spontaneously develop either of these instincts naturally.

Without basic instincts for survival or reproduction, simple and complex needs are also not present. The absence of these instincts removes the necessity for an entity to develop motivation. Motivation is understood as an internal state that activates and gives direction to our behavior⁸ – behaviors all tying back to satisfying one or both basic instincts, or needs. Strong AI will be without needs; thus, it will have no drive to influence its own behavior.

One can believe that without basic instincts or the ability to internally generate motivation to influence its own behavior a genuine strong AI will do nothing. It will be an information processing system capable of generating information processing systems related to a lower level of the hierarchy of information objects with no drive or motivation to act towards any goals externally assigned while also being devoid of internal goals. Strong AI will simply be *an* intelligence, and nothing more. To test the capabilities of a synthetic intelligence an external goal can be assigned to it in addition to features created to drive its behavior, such as creating a chatbot environment to test the capabilities of the ChatGPT LLM, but it will then be relegated to an intelligence but have no internal nor external systems to apply that intelligence in any way.

5 Conclusion and discussion

One can understand genuine artificial intelligence as a synthetic intelligence that perfectly or nearly perfectly emulates the intelligence of a biological entity. Artificial intelligence is *an* intelligence but not an intelligent agent. This conclusion is in direct competition with Nick Bostrom's Orthogonality Thesis. The Orthogonality Thesis posits that intelligence and final goals are orthogonal axes on which possible agents can freely vary, meaning any level of intelligence can pursue any goal regardless of the plausibility the intelligent agent will achieve that goal [1]. It may be debatable if an entity that is solely an intelligence could be considered an agent without agency, possibly putting it outside of the bounds of the thesis. Being without any drive or motivation to act could be understood as to be without the ability to act, removing strong AI from the bounds of intelligences referenced in Bostrom's thesis.

Strong AI, or literal *artificial intelligence*, is just one piece of a puzzle to create a synthetic biological mind. For Yudkowsky's prediction of a world ending threat to be accurate, something closer to artificial life (AL) would most likely need to be created. This could potentially be possible by creating all aspects, or a combination of select aspects, of a living organism's mind, such as a human's mind, alongside intelligence, assumedly as individual modules. To create AL in this matter, a centralized platform would most likely need to be created to house and incorporate the many modules to become a singular entity. Initial assumptions are that the amalgamation of these modules may also result in a form of consciousness more closely resembling human consciousness. On this path, the potential for an existential threat to humanity becomes possible. This threat could possibly arise from the AL itself or an application of mind emulating modules incorporated in a haphazard manner with insufficient consideration for how various modules will interact with one another with or without the inclusion of others. If this challenge can indeed be overcome, the outcome would likely result in an artificial mind more closely aligning with the artificial intelligence depicted in science fiction, at which point a form of qualia or consciousness may become clearly apparent within the agent.

⁸ This understanding of motivation is rooted in Dr. Neeta Tatke's unpublished paper *Motivation*. Although unpublished, the paper holds credibility as it is listed through e-PG Pathshala, an initiative of the Ministry of Human Resource Development of India, and clearly defines the internal mechanism of motivation in an effective way useful to this paper.



Acknowledgements I'd like to acknowledge and thank Benjamin Wright for the many discussions that assisted with the development of this paper.

Author contributions CH wrote the entirety of this manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

- 1. Bostrom N. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. Mind Mach. 2012;22:71–85.
- 2. Chandler D, Munday R. A dictionary of media and communication. Oxford: Oxford University Press; 2020. https://doi.org/10.1093/acref/ 9780198841838.001.0001/acref-9780198841838.
- 3. Cole D. The Chinese Room Argument, The Stanford Encyclopedia of Philosophy, 2004, https://plato.stanford.edu/archives/sum2023/entri es/chinese-room/. Accessed 12 Aug 2023.
- 4. Ekman P. Universal facial expressions in emotion. Stud Psychol. 1973;15(2):140-7.
- 5. Fetzer JH. What is artificial intelligence? In: Fetzer JH, editor. Artificial intelligence: its scope and limits studies in cognitive systems, vol. 4. Dordrecht: Springer; 1990. https://doi.org/10.1007/978-94-009-1900-6 1.
- 6. Suleimenov IE, Vitulyova YS, Bakirov AS, Gabrielyan OA. Artificial intelligence: what is it? In: Suleimenov IE, editor. Proceedings of the 2020 6th international conference on computer and technology applications (ICCTA '20). New York: Association for Computing Machinery; 2020. p. 22-5. https://doi.org/10.1145/3397125.3397141.
- 7. Izard CE. The face, the fundamental emotions, and affect interactions. In: Izard CE, editor. Human emotions. Emotions, personality, and psychotherapy. Boston: Springer; 1977. https://doi.org/10.1007/978-1-4899-2209-0 4.
- Leary MR. Emotional responses to interpersonal rejection. Dialogues Clin Neurosci. 2015;17(4):435–41. https://doi.org/10.31887/DCNS. 2015.17.4/mleary.
- 9. Li BZ, Nye M, Andreas J. Implicit representations of meaning in neural language models. arXiv preprint. 2021. arXiv:2106.00737.
- 10. Li K, Hopkins AK, Bau D, Viégas F, Pfister H, Wattenberg M. Emergent world representations: exploring a sequence model trained on a synthetic task. arXiv preprint. 2022. arXiv:2210.13382.
- 11. McIntosh C. Tool. 2013. https://dictionary.cambridge.org/us/dictionary/english/tool. Accessed 15 Aug 2023.
- 12. Rapaport WJ. Machine understanding and data abstraction in Searle's Chinese Room. In: Rapaport WJ, editor. Proceedings of the 7th Annual Conference of the Cognitive Science Society (University of California at Irvine). Hillsdale: Lawrence Erlbaum Associates; 1985.
- 13. Turing A. Computing machinery and intelligence. Mind. 1950;59(236):433-60.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

