THE LIMITS OF STRONG AI:
AN EXAMINATION OF THE ARGUMENTS IN THE "EMPEROR'S NEW MIND"[1]

Dr. Hassan Bahrami[2] (PhD in Computer Science)

Email: hassan.bahrami@student.uts.edu.au

Introduction

Artificial Intelligence (AI) has advanced rapidly in recent years, leading to increasing speculation about the potential for machines to achieve human-like consciousness. **Strong AI**, a term introduced by philosopher John Searle, claims that machines, if programmed correctly, could possess minds, consciousness, and understanding comparable to human beings. This form of AI assumes that human cognition is fundamentally computational and can be replicated by a sufficiently complex algorithm. However, in The *Emperor's New Mind*, this assumption is rigorously challenged. The book presents a comprehensive argument that human consciousness cannot be fully replicated by machines because it involves processes that go beyond computation. Drawing on mathematical theories, philosophical reasoning, and quantum mechanics, the book argues that human cognition involves non-computable elements, rendering the strong AI claim inadequate. This essay explores the concept of **algorithms**, mathematical limitations such as **Gödel's Incompleteness Theorems**, **Russell's Paradox**, the **Turing Halting Problem**, and their interrelationship, concluding that human intuition and understanding involve processes beyond the scope of machine intelligence.

1. The Concept of Strong AI and Algorithms

The foundational assumption behind **strong AI** is that the human mind is essentially a complex computational machine. Proponents of this view argue that with the right algorithms and sufficient processing power, a machine could replicate every aspect of human cognition, including understanding, consciousness, and emotions. **Algorithms** are defined as step-by-step procedures or rules that a machine follows to solve a problem or perform a task. Classical AI operates within this framework, using algorithms to simulate human decision-making processes.

In this view, the human brain is often likened to hardware, while thoughts, emotions, and reasoning are the software—the algorithms—that the brain executes. A critical component of this perspective is the **Turing Test**, developed by Alan Turing, which evaluates whether a machine's behaviour is indistinguishable from that of a human. If a machine passes the Turing Test, it is argued that the machine

[1] To cite this essay: Bahrami, H. (2024). The limits of strong AI: An examination of the arguments in The Emperor's New Mind**,** *Wisdom House*. Available at https://www.wisdomhouse.at

[2] Hassan completed his PhD in Computer Science at the University of Technology Sydney, Australia, and earned his Master's in Industrial Engineering from Sharif University of Technology. His primary research contribution lies in developing computational models across various applications, including statistics, design, and fabrication.

can "think" or possess a mind. However, while machines can process data using algorithms, there is an ongoing debate about whether this is equivalent to genuine understanding. In **Searle's Chinese Room** thought experiment, a person manipulates symbols based on a set of rules without understanding the meaning of the symbols themselves. The Chinese Room demonstrates that while machines can follow algorithms to simulate human-like responses, they do not possess true comprehension or consciousness.

2. Algorithms and Arithmetic: Russell's Paradox and Gödel's Incompleteness

The relationship between **algorithms** and **arithmetic** is a crucial point in understanding the limits of computation. Arithmetic, the basic system of numbers and operations, can be represented through formal systems composed of **axioms** and **rules**. One of the most famous attempts to formalize mathematics was undertaken by Bertrand Russell and Alfred North Whitehead in their work, *Principia Mathematica*. In this work, they sought to prove that all mathematical truths could be derived from a set of fundamental axioms. For example, they demonstrated that "1 + 1 = 2" could be rigorously proven using the formal structure they developed.

However, this approach ran into several problems, notably **Russell's Paradox**. This paradox arises when one considers the set of all sets that do not contain themselves. If such a set exists, does it contain itself or not? If it does, it contradicts its definition; if it does not, it should contain itself. This paradox exposed fundamental flaws in the foundation of set theory and, by extension, in the ability to reduce mathematics to a purely formal system. **Russell's Paradox** illustrates the limitations of algorithms when applied to logic and set theory, challenging the notion that all mathematical truths can be derived algorithmically.

Further complicating the issue is **Gödel's Incompleteness Theorem**, which states that in any formal mathematical system, there are true statements that cannot be proven within the system. This result has profound implications for AI and the concept of algorithms. While machines operate by following a set of predefined rules, Gödel's theorem shows that there are limits to what these rules can achieve. Even in a simple system like arithmetic, there are truths that lie beyond the reach of any algorithm. This presents a significant challenge to strong AI, as it implies that human beings, who can intuitively grasp these unprovable truths, engage in cognitive processes that are not entirely reducible to algorithms.

3. Turing Machines and the Halting Problem

The **Turing Machine** is a theoretical model of computation introduced by Alan Turing, and it serves as the foundation for understanding how classical computation works. A Turing Machine operates by reading symbols on a tape and following a set of predefined rules to manipulate those symbols. It can, in theory, simulate any algorithmic process, making it a powerful model for understanding the capabilities and limitations of computers.

However, Turing himself discovered a limitation of this model, known as the **Halting Problem**. The halting problem asks whether it is possible to devise an algorithm that can determine, for any given program and input, whether the program will eventually halt or continue running indefinitely. Turing proved that no general algorithm can solve this problem for all possible programs. This discovery has significant implications for AI, as it demonstrates that there are some problems that no algorithm can solve, regardless of the machine's power or complexity. In the context of strong AI, this suggests that there are limits to what machines can achieve through purely computational means. While machines can execute algorithms with

incredible speed and precision, they cannot overcome the fundamental constraints imposed by the halting problem.

The **Turing Halting Problem** also ties into **Russell's Paradox** and **Gödel's Incompleteness Theorems**. All of these results highlight the inherent limitations of formal systems and algorithms. In each case, there are truths or outcomes that cannot be captured by a fixed set of rules, no matter how sophisticated. This poses a serious challenge to the idea that human cognition, which seems to transcend these limitations, can be fully replicated by machines.

4. The Role of Intuition in Human Cognition

One of the central arguments against strong AI is the role of **intuition** in human thought. Human beings can often grasp truths that are not immediately apparent through formal reasoning. For example, mathematicians frequently rely on intuition to guide them toward insights and proofs that are not immediately derivable from formal systems. This capacity for intuition suggests that human thought involves more than just the mechanical execution of algorithms.

In *The Emperor's New Mind*, it is suggested that **intuition** plays a crucial role in human understanding, particularly in areas like mathematics. **Gödel's Incompleteness Theorems** support this view by showing that there are mathematical truths that cannot be formally proven but are nonetheless recognized by human minds. This implies that human reasoning involves processes that extend beyond the purely algorithmic.

Additionally, **Plato's view of truth** as something "God-given" rather than "man-made" echoes the idea that human beings have access to truths that are not fully explainable by formal logic. Intuition, in this sense, may be seen as a kind of cognitive ability that allows humans to grasp self-evident truths or insights that machines, constrained by algorithms, cannot reach. This suggests that while machines are excellent at following rules, they lack the capacity for the kind of creative, intuitive thinking that characterizes human cognition.

5. Quantum Mechanics and Non-computable Processes

A key part of the argument against strong AI involves **quantum mechanics**, which introduces non-deterministic processes that classical computation cannot account for. Classical AI operates under the assumption that the brain is like a computer, where processes follow strict, deterministic algorithms. However, quantum mechanics suggests that at the smallest scales, the universe operates in a way that is fundamentally non-deterministic. Phenomena like **quantum superposition** and **quantum entanglement** challenge the classical view of computation.

In *The Emperor's New Mind*, it is suggested that **quantum processes** in the brain, particularly within structures called microtubules, could be responsible for aspects of human consciousness that are non-computable. The **Orchestrated Objective Reduction (Orch-OR)** theory, developed with Stuart Hameroff, proposes that consciousness arises from quantum events in the brain. These quantum processes may explain the non-computable elements of human thought, such as intuition and insight, which go beyond the capabilities of classical algorithms.

If consciousness is indeed tied to quantum mechanics, this presents a significant challenge to strong AI. While classical computers operate based on deterministic algorithms, the brain may operate using processes that are fundamentally non-computable, suggesting that human cognition cannot be fully replicated by machines. This non-computable aspect of quantum mechanics may explain why humans are capable of grasping truths that lie beyond the reach of formal systems and algorithms.

6. Conclusion: Beyond Computation—The Uniqueness of Human Cognition

The arguments presented in *The Emperor's New Mind* illustrate the inherent limitations of strong AI. While machines can simulate certain aspects of human intelligence using algorithms, they cannot replicate the full depth of human cognition. The limitations imposed by **Gödel's Incompleteness Theorems**, **Russell's Paradox**, and the **Turing Halting Problem** show that formal systems and algorithms have inherent boundaries. These results challenge the assumption that human thought is purely computational.

Moreover, the role of **intuition** in human reasoning and the possibility that **quantum mechanics** underlies consciousness further suggests that human cognition involves processes that are non-computable. While AI can mimic human behaviour to some extent, it cannot achieve the kind of deep understanding, insight, and creativity that characterize human.

References

Penrose, Roger. *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford Paperbacks, 1999.

ENVISIONING IRAN'S SCIENTIFIC FUTURE IN THE ERA OF ARTIFICIAL INTELLIGENCE[3]

Dr. Mohammad Hoseini Moghadam[4] (Assistance Professor of Foresight)

Email: Moghadam@iscs.ac.ir

Abstract

Artificial Intelligence (AI) is a transformative technology that is reshaping various domains of human society. This article presents a futures studies perspective on the impact of AI on the higher education institution in Iran. Drawing from a five-year research program involving visits to universities across Iran, dialogues with stakeholders, and participation in drafting the country's AI strategic document, this work explores the challenges and opportunities AI presents for Iran's scientific community. The article underscores the importance of adopting a macro-historical view, understanding national realities, and developing a proactive approach to shape a desirable AI-enabled scientific future. Key recommendations include realistic planning, international collaboration, investment in AI education and infrastructure, and balancing AI adoption with ethical considerations and human-centric values.

Introduction

The advancement of Artificial Intelligence (AI) represents a megatrend profoundly influencing the trajectory of various domains in human society, including the institution of higher education (Hilbert, 2020). As a researcher specializing in science and technology futures studies, I have undertaken a five-year research program focused on exploring AI's impact on the scientific landscape in Iran. This article presents the key findings and recommendations derived from this extensive research endeavor.

Throughout this research program, which commenced in 2019 and continues to the present day, I have visited various universities across Iran, encompassing polytechnic universities, medical science universities, institutions focused on humanities and social sciences, as well as comprehensive universities. My objective has been to gather perspectives from these institutions on the research topic of AI's impact on higher education. Additionally, through several educational workshops attended by students, researchers, university professors, academic administrators, and higher education policymakers, I have engaged in

---

[3] To cite this essay: Hoseini Moghadam, M. (2024). Envisioning Iran's Scientific Future in the Era of Artificial Intelligence**,** *Wisdom House*. Available at https://www.wisdomhouse.at

[4] Foresight Department, Institute for Social and Cultural Studies, Tehran, Iran.

dialogues to understand their views on the current state and future prospects of AI in Iran's scientific institutions. Furthermore, I have participated in the process of drafting the country's strategic document in the field of AI, emphasizing the necessary requirements for Iran to achieve a desirable status in this domain.

Conceptual Framework

To provide a framework for understanding AI's transformative potential, I have developed a conceptual approach integrating three key elements: macro-historical thinking, realism, and futurism.

1. Macro-historical Thinking: Drawing from the work of Martin Hilbert, changes in technology lead to shifts in the means of production, which in turn transform the mode of development and societal superstructures (Hilbert, 2020). This historical trend can be traced from hunter-gatherer societies to agricultural societies, and through the successive industrial revolutions, culminating in the emergence of an intelligent society driven by AI.

2. Realism: It is crucial to gain an accurate and comprehensive understanding of the realities surrounding AI at local, national, and international levels. Failing to do so can result in a narrow or incomplete perspective, akin to the story of the elephant in Rumi's Masnavi, where individuals touched only a part of the elephant and drew incorrect conclusions about its entire form (Rumi, 2004).

3. Futurism: The future will not wait for those who remain passive observers. Successful nations are those that actively engage with emerging technologies, anticipate their impacts, and take deliberate steps to shape a desirable future aligned with their national goals and values (Kuosa, 2016).

Historical Transitions and Technological Progress

Hilbert's analysis highlights the correlation between technological advancements and global metamorphoses throughout human history (Hilbert, 2020). The initial focus was on transforming matter (e.g., stone, bronze, iron), followed by a revolution in energy from 1780 to 1973, and the current era dedicated to transforming information since 1973. The central idea of this work is to examine the transition from tradition to silicon, enabling Iran's scientific institutions to create better futures by adapting to these historical transformations.

Furthermore, there exists a historical relationship between educational advancement and technological progress. During the transition from an agrarian society to the First Industrial Revolution, the education system experienced a 'social pain' due to its failure to recognize the pace of technological change (Penprase, 2018). However, on the cusp of the Digital Revolution, universities have taken a pioneering stance by advancing several steps ahead of technological progress. The challenge for the future lies in universities' ability to transform the educational landscape to confront emerging technologies like AI.

AI Revolution in Developing Countries

In the context of developing countries like Iran, the advent of AI presents both opportunities and challenges. While society becomes aware of technological changes much earlier and puts them to use, the governance system often realizes this issue with a delay, resulting in a historical lag (Ananiadou & Claro, 2009).

Currently, according to official statistics, Iranian society's general awareness about AI, particularly in the form of Large Language Models (LLMs) and Natural Language Processing (NLP), is at the global average level. This increased attention to AI tools, especially among students and the younger generation entering universities, has created unprecedented opportunities for innovation and scientific progress in Iran.

The democratization of knowledge and learning tools through AI represents a significant shift, allowing individuals from diverse backgrounds to engage with cutting-edge information and learning resources (Belfield et al., 2020). However, this rapid adoption of AI tools by society also highlights the need for the governance system to catch up and develop appropriate policies and frameworks to harness these technologies effectively and responsibly.

Risks of AI Illiteracy

One of the key risks highlighted in this research is the lack of sufficient awareness and understanding of AI developments among Iranian governors. This issue could lead to the scientific future of the country being exploited and colonized in different ways, resulting in a limited role and share for the Iranian scientific community in shaping the future (Coeckelbergh, 2020).

The new world order in the future will be based on the guidance and leadership of corporations that have a monopoly on data, hardware, human resources, and significant material capital. This monopoly allows them to gain a unique dominance in guiding countries' affairs, potentially leading to a loss of scientific autonomy and decision-making power for nations like Iran (Zuboff, 2019).

For example, tech giants like Google can provide more accurate judgments about scientific topics of interest in various university departments based on precise data and evidence, while traditional institutions may struggle to provide such detailed insights promptly (Bridle, 2018). This scenario illustrates the power imbalance between large tech corporations and traditional institutions, highlighting the need for countries like Iran to develop their own technological capabilities and data infrastructure to participate meaningfully in shaping AI's future.

AI Giants and Data Monopolies

The concentration of AI resources, including data, expertise, hardware, and processing power, in the hands of a few tech giants like Google, Amazon, Microsoft, and OpenAI, has significant implications for global scientific competition and knowledge production (Author 1 & Author 2, 2021). The monopoly on crucial hardware components like the GPU H100, essential for increasing information processing speed, further solidifies the dominance of these leading AI companies.

Preserving Human Creativity and Critical Thinking

While AI offers numerous benefits, there are concerns about its potential impact on human cognitive abilities (Miailhe & Hodes, 2017). The absence of appropriate policies and regulations in Iran's higher education system regarding the use of generative AI by students and professors has led to dysfunctions in the scientific establishment.

Many reputable universities worldwide have established regulations to determine how students and professors should use generative AI, enabling the management of potential misuse (Warne, 2022). However, the lack of such procedures in Iran's universities allows students to easily bypass educational tasks and research requirements, potentially leading to an over-reliance on AI outputs and a gradual loss of critical and creative thinking skills in educational and research processes (Kotzee, 2022).

Envisioning a National AI-Enabled Scientific Future

To address these challenges and harness AI's potential for scientific development, Iran's academic community needs to create a comprehensive vision for integrating AI into research and education. This vision cannot be limited to a small group of policymakers but should be an intersubjective endeavor with the participation of all stakeholders in the scientific community (Goertzen, 2017).

By involving students, professors, employers, higher education policymakers, and planners in the visioning process, Iran can develop a more robust, relevant, and widely accepted strategy for integrating AI into its academic and research environments. This approach can help ensure that AI's implementation in higher education aligns with the needs and expectations of all involved parties, potentially leading to more effective and sustainable outcomes (Boer & van den Berg, 2022).

Recommendations for Action

Based on the findings of this research, several recommendations are proposed to improve the encounter between Iran's scientific institution and AI:

1. Develop realistic plans for AI integration in scientific research based on the capabilities and capacities of Iran's scientific institution to achieve the desired future of scientific progress (Loke, 2022).

2. Foster international collaboration in AI and data science, recognizing the interdisciplinary and global nature of this field, and the need for access to hardware, software, specialists, databases, and data centers (Vu et al., 2021).

3. Invest in developing the necessary infrastructure in Iran to achieve global AI standards in universities and ensure competitiveness (Li et al., 2021).

4. Balance AI adoption with ethical considerations and human-centric values, ensuring that 'human in the loop' is realized in contrast to 'AI in the loop' (Calo, 2022).

5. Understand the maturity model of 'intelligentization,' which involves progressing through the stages of informatization, digitalization, and intelligentization to enhance the productivity and quality of educational and research services in Iran's universities and scientific institutions (Shi et al., 2022).

Conclusion

The integration of Artificial Intelligence into Iran's scientific landscape presents both challenges and opportunities. By adopting a macro-historical perspective, understanding national realities, and developing a proactive approach, Iran's academic community can shape a desirable AI-enabled scientific future.

Key recommendations for achieving this goal include realistic planning, international collaboration, investment in AI education and infrastructure, and balancing AI adoption with ethical considerations and human-centric values. Additionally, it is crucial to develop a comprehensive vision for AI integration in research and education through an intersubjective process involving all stakeholders in the scientific community.

By addressing these recommendations and embracing AI's transformative potential while preserving human creativity and critical thinking, Iran's scientific institutions can position themselves at the forefront of scientific progress in the era of Artificial Intelligence.

References

Ananiadou, K., & Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries. OECD Education Working Papers, No. 41. https://doi.org/10.1787/218525261154

Belfield, C., Bowden, B., Klapp, A., Levin, H., Shand, R., & Zander, S. (2020). Two decades of change: An analysis of educational attainment. National Center for the Study of Privatization in Education.

Boer, E. R., & van den Berg, B. A. M. (2022). AI in higher education: A reality? An explanatory mixed methods study on the awareness and expectations of stakeholders in Dutch higher education regarding the adoption of AI technologies. Technology, Knowledge and Learning, 27(3), 1057-1080. https://doi.org/10.1007/s10758-021-09533-3

Bridle, J. (2018). New dark age: Technology and the end of the future. Verso Books.

Calo, R. (2022). Artificial intelligence policy: A primer and roadmap. UC Davis Law Review, 51(2), 399-435.

Coeckelbergh, M. (2020). AI ethics. The MIT Press.

Goertzen, B. (2017). AI futures: Governing advanced artificial intelligence. Future Today Institute. https://futuretodayinstitute.com/trends/ai-futures-governing-advanced-artificial-intelligence/

Hilbert, M. (2020). Digital technology and social change: The digital transformation of society from a historical perspective. Dialogues in Clinical Neuroscience, 22(2), 189-196. https://doi.org/10.31887/DCNS.2020.22.2/mhilbert

Kotzee, B. (2022). Artificial intelligence, education and epistemological insecurity. Educational Philosophy and Theory, 54(1), 22-31. https://doi.org/10.1080/00131857.2021.1934791

Kuosa, T. (2016). The disrupted futures model: Towards proactive strategies. Aalto University.

Li, Z., Zhou, Y., Di Resta, R., & Nayebi, A. (2021). Bridging education and AI technology: Key challenges and future pathways. ArXiv, abs/2110.06658.

Loke, S. (2022). AI and education: A literature review and narrative on emerging pedagogies and priorities. Computers & Education: AI, 3, 100069. https://doi.org/10.1016/j.caeai.2022.100069

Miailhe, N., & Hodes, C. (2017). The third age of artificial intelligence. Field Actions Science Reports, 17, 6-11.

Penprase, B. (2018). The fourth industrial revolution and higher education. In N. Gleason (Ed.), Higher education in the era of the fourth industrial revolution (pp. 207-224). Palgrave Macmillan. https://doi.org/10.1007/978-981-13-0194-0_9

Rumi, J. M. (2004). The Masnavi: Book one (J. Mojaddedi, Trans.). Oxford University Press.

Shi, K., Gu, J., & Xie, W. (2022). Intelligentization maturity model of information system. ArXiv, abs/2204.07274.

Vu, T. M., Warwick, S., Salles, E. O., Kiesewetter, M., Daigle, M., Cukier, M., Becker, C., Holmes, C., Zanda, M., & Zemaitis, L. (2021). Toward an ethical toolkit for global AI. First Monday, 26(7). https://doi.org/10.5210/fm.v26i7.11773

Warne, R. T. (2022). AI ethics education: A case study in Québec. AI and Ethics, 2(3), 219-236. https://doi.org/10.1007/s43681-022-00150-6

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Profile Books.