FUTURITY PHILOSOPHY

DOI: https://doi.org/10.57125/FP.2023.12.30.03

How to cite: Namestiuk, S. (2023). Challenging the Boundary between Self-Awareness and Self-Consciousness in AI from the Perspectives of Philosophy. *Futurity Philosophy*, *2*(4), 43-60. https://doi.org/10.57125/FP.2023.12.30.03

Challenging the Boundary between Self-Awareness and Self-Consciousness in AI from the Perspectives of Philosophy

Svitlana Namestiuk

PhD in Philology, Associate Professor Department of Foreign Languages, Bukovinian State Medical University, Chernivtsi, Ukraine, https://orcid.org/0000-0002-9427-5675

***Corresponding author**: *Lapetitelarousse83@gmail.com.*

Received: September 17, 2023 | **Accepted:** November 28, 2023 | **Published:** December 30, 2023

Abstract: With the progression of technology, scientists are dealing with the challenges brought about by the creation of smart AI models designed for actual use. This paper aimed to explore the complex boundary between consciousness and self-awareness in artificial intelligence (AI) systems from a philosophical perspective. Methods - The literature review conducted as part of this study was based on materials obtained from such databases as WOS, Web of Science, Scopus, and Google Scholar. The findings of the study indicated that, in the present day, philosophy is analysing how technological advancements have led to a focus on conceptual frameworks that define consciousness as the capacity of artificial intelligence to identify its functional states. Self-awareness is also being viewed as the ability to reflect and recognise oneself within a larger context. Drawing on the principles of symbolic AI and connectionism, the study considers whether AI can have a form of self-awareness similar to biological entities and how this relates to philosophical debates about consciousness. The study delved into the symbolic representation of knowledge and logical reasoning as tools for AI to achieve a primary form of self-awareness. This is contrasted with a connectionist approach that models the emergent properties of neural networks, suggesting a pathway to more complex forms of self-awareness. The article discussed the limitations and potential of these approaches, including the ethical implications of creating systems that can mimic aspects of human self-awareness and self-concept. By thoroughly examining the current AI technologies and their underlying philosophical principles, the article aims to

give a detailed look at the ongoing debate on the cognitive abilities of AI and whether it can move beyond basic self-awareness to a more profound level of self-awareness. The conclusion is that although AI may exhibit behaviours indicative of self-awareness, the subjective experience of self-awareness remains a uniquely human trait, which calls into question the boundary between human and machine cognition.

Keywords: consciousness, subjective experience, cognition, possibility of autonomous choice, principle of consciousness, programme.

Introduction

Philosophers of technology have always sought to uncover the deeper nature of technology and its relationship to social structures and human experience. During the first half of the twentieth century, their research frequently exposed a disconnect between advancements in technology and the individual components of human existence. In today's world, technology is often perceived as an independent factor that has a destructive impact on key aspects of human nature. Philosophers who focus on the historical and universal aspects of technology sometimes distance themselves from the real impact of specific technological developments (Pelau, Dabija & Ene, 2021). In his study, Nikolenko (2022) analyses the limitations of artificial intelligence, especially in the context of philosophy. The author argues that despite AI's ability to mimic some forms of human intelligence and its high data processing speed, it cannot fully replace human philosophical thinking. The author provides convincing evidence in favour of this position. At the same time, Kampen, Romanchuk & Palij (2022) states that artificial intelligence, although progressive, does not have the capacity for consciousness or subjective experience, which are key to philosophical thought. They argue that philosophy involves not only logical reasoning, but also a deep reflection on existential and ethical issues.

In the current discourse, technology is perceived as an autonomous and deterministic entity that has the potential to eradicate essential aspects of human existence. Philosophers have tended to oversimplify the complex interplay between technology and humanity by reducing it to abstract historical or transcendental theories, thereby overlooking the nuanced and concrete impacts of technological developments on society.

Heflin (2020) emphasises that the basis of philosophy is the capacity for doubt, questioning, and independent thinking. These characteristics are unique to humans and cannot be replicated in artificial intelligence. Can machines comprehend the diversity and depth of human thinking, which is characterised by subtle shades and introspection? It is important to recall the importance of Socratic Mayaevtica, a method of dialogue that stimulates personal and deep reflection. Can a machine really participate in such a process? The answer is obviously no, because such a depth of reflection remains beyond the reach of modern AI. Some scholars have strongly criticised predictions that artificial intelligence is being capable of developing consciousness or independent thinking. They point out that such ideas are often overly optimistic by AI advocates (Craswell, Mitra, Yilmaz, Campos & Voorhees, 2020). But is there any reason to consider that machines will be able to think autonomously in the future? Given the current and fundamental limitations of technology, this scenario seems impossible. Are machines capable of surpassing the parameters of their programming and achieving a state of consciousness? At present, hopes for such a development of artificial intelligence do not have a sufficient scientific basis.

In contemporary scholarship, a novel empirical paradigm has surfaced in the realm of technological inquiry, informed by the philosophical perspectives of scholars who consider the cognitive repercussions of specific technologies. The issue of distinguishing between human self-consciousness

and the possible self-consciousness of artificial intelligence is key to such considerations. Research that focuses on the interaction between technology and society highlights this interdependence. This "empirical turn" contributes to an understanding of the universality of technology and the multiple roles it plays in the social context. In this aspect, the research aims to define consciousness and self-awareness, to establish clear definitions that apply to both humans and artificial intelligence, based on philosophical theories and cognitive science.

Research Problem

The study sought to explore the concept of artificial intelligence possessing consciousness and self-awareness, and to examine how these qualities can be distinguished and measured. The research aimed to contribute to the ongoing debate in the field of artificial intelligence about the ethical and moral implications of creating machines that may possess self-awareness. The study of classical and contemporary philosophical theories of consciousness and self-awareness will help to reveal the views of modern philosophers who have studied the nature of consciousness and its consequences.

Research Focus

This paper aimed to draw on the knowledge of cognitive science, neuroscience, and psychology in order to understand how human self-awareness is theorised and empirically studied. The paper aims to define consciousness and self-awareness in order to establish clear boundaries between these concepts in human and artificial intelligence. This implied revealing the meaning of these terms from the philosophical, psychological, and technical points of view. An analysis of the shift towards empirical methods in philosophical research, especially in areas that intersected with technology and society, could contribute to the understanding of the limits and nuances of artificial intelligence's self-awareness. The chosen objectives allowed the presented research to make a significant contribution to the ongoing discourse on the philosophical implications of artificial intelligence technologies, especially with the regard to its potential for consciousness or self-awareness. This interdisciplinary approach provided a comprehensive study of the complex relationships between technology and philosophical reflection.

Research Aim and Research Questions

The aim of this paper was to study philosophical discussions about consciousness and self-awareness, as well as the application of these concepts to the development of AI.

The purpose of the article was to assess how modern theories of consciousness can be applied to AI.

To determine whether the concepts of self-awareness and self-awareness can be attributes of AI.

To analyse ethical dilemmas related to AI with advanced cognitive abilities.

To study philosophical arguments about the possibility of true AI self-awareness.

Research questions:

1.Which modern theories of consciousness are most relevant for understanding cognitive processes in AI?

2.Can AI have "consciousness" in the traditional sense of the term?

3.What ethical issues arise when creating AI with a high level of self-awareness?

This research can contribute to a broader discourse on AI and its place in our understanding of consciousness and self-awareness. It is interdisciplinary, engaging philosophy, cognitive science, AI development, and ethics.

Literature Review

Artificial intelligence (AI) is at the forefront of modern technological advances, prompting deep philosophical reflections on its growing capabilities and implications for humanity. As we continue to explore the intricacies of AI cognitive processes and self-awareness, inquiries emerge regarding the underlying mechanisms of consciousness and its potential interfaces within non-organic entities. This research not only challenges traditional notions of identity, but also calls us to rethink basic philosophical questions about what it means to be self-aware and to feel. Assunção, Patrão, Castelo-Branco, and Menezes (2022) analyse the goals and methodology of artificial intelligence through the lens of the rationalist understanding of intelligence that has been espoused by many rationalist philosophers throughout history. They point out that humans, when acquiring knowledge in a particular field, initially rely on formal rules, but eventually move on to using empirical rules and intuition. A rationalist approach to understanding AI is reflected in the concept of symbolic AI, where intellectual processes are seen as processing information represented by symbols, limiting intelligence to manipulating symbols. Ng & Leung (2020) considered the approach to artificial intelligence as a synthesis of three main theses. The psychological thesis states that the human mind works on the basis of symbolic manipulation. The epistemological thesis assumes that knowledge can be fully formalised. The ontological thesis assumes that the structure of reality can be expressed by formalisation.

Adherents of Dreyfus's theory not only criticised these theses, but also proposed alternative concepts that they considered key to understanding intelligence. Particularly, Dreyfus believed that intelligence is embodied and contextualised, which means that it is dependent on the physical body and a specific environment (Brinkman & Grudin, 2023). This view emphasises the importance of real-world context and relationships, rejecting the idea that reality can be reduced to a set of formalised rules, and questions the ability of symbolic-manipulative machines to function effectively outside of strictly defined formal systems.

Bajohr (2020) is optimistic about the connectionist approach to artificial intelligence, which is based on the idea that intelligent behaviour can emerge from networks similar to the neural connections of the human brain. However, the author expresses doubts about the possibility of reproducing the complexity of the human brain in machines. His work has provoked widespread debate and forced some researchers to reconsider their views, leading to the development of systems that better match his concepts.

Modern researchers have expressed criticism of rigid symbolic methods in the field of artificial intelligence (Chun & Elkins, 2022). Over the past few decades, attempts have been made to develop hybrid intelligent systems that integrate non-symbolic approaches within the framework of symbolic AI.

The concept of free will is the subject of intense philosophical debate (Barron, 2023). It requires a thorough analysis of human cognition and the possibility of autonomous choice. Despite the diversity of approaches to understanding the human mind, the existence of free will is a matter of considerable controversy (Dodda, Maruthi, Yellu, Thuniki & Reddy, 2021). and unsound in a deterministic universe. From a scientific standpoint, this debate extends into the realms of neuroscience and psychology, where studies on decision-making processes suggest that our actions are influenced by a complex interplay of biological, environmental, and genetic factors. In light of such evidence, it becomes increasingly difficult to uphold the notion of free will as a universally applicable concept. Therefore, the scientific analysis of

human behaviour leans towards a deterministic explanation, with the acknowledgment that our perceived choices are ultimately predetermined by a web of causal factors beyond our immediate control (Wei, 2022). Such a denial of free will can be seen as a pragmatic contradiction, since the rejection of free will inevitably makes any such statement self-destructive.

According to Haikonen (2020), the issue of free will remains important because scientific theories propose that all events occur in accordance with the laws of nature. This presents us with a choice: either to attribute unique properties to humans or to reject the determinism of the laws of nature to avoid being recognised as deterministic biological mechanisms (Edelman, 2023). The first alternative correlates with numerous philosophical doctrines, including Descartes' dualistic theory, which divides the world into matter and spirit integrated in humans. The second approach offers a more holistic understanding, using modern advances in physics, such as relativity and quantum mechanics, to demonstrate that our free will can be based on the non-determinism of natural processes (Zhao, Li & Xu, 2022).

Descartes' dualistic perspective, which is supported by other philosophers, denies the existence of free will in anything other than humans, reducing interest in discussions about free will and AI (Esmaeilzadeh & Vaezi, 2021). Instead, the holistic approach is better suited to such discussions, although it leads to conclusions that depend on the physical assumptions underlying the attribution of free will to humans or machines. This may be useful in a philosophical context, but has limited application in the field of computer science.

In their work, Agarwal & Edelman (2020) note that it is also possible to recognise the dialectical nature of humanity, which includes both deterministic and free aspects that matter. This approach allows us to see free will as a transcendent feature of human beings, not limited to physical circumstances. However, such a transcendental vision cannot be applied to the analysis of free will in artificial intelligence, as transcendental conditions cannot be modelled or visualised.

This paper attempted to reformulate the complex idea of consciousness and self-knowledge into a concept that can be used to analyse artificial intelligence, in accordance with the empirical approach in the philosophy of technology. Therefore, the paper focused on the social role of humans and AI.

In this context, the paper proposed two fundamentally different methods of research in this area. The first approach delved into a deep philosophical reflections on the nature of consciousness and self-knowledge and the human ability to resist the “laws” of nature - the physical aspect. In the context of articles about artificial intelligence, this leads to a philosophical debate that is more about human nature than about the nature of computers, as we find ourselves in a situation where we have to defend our own will and consciousness to be able to say anything about it. As a result, the debate turned into a comparative analysis between humans and computers, where both sides fail to recognise each other.

Another method of analysis emphasised that it is impossible ignoring the existence of our consciousness and self-knowledge. As noted earlier, ignoring this fact is meaningless. Every social dilemma requires us to make choices, and in the context of these choices, the concept of consciousness becomes key. The main idea is that consciousness is a fundamental prerequisite for the functioning of our society, regardless of its physical validity - it is a social dimension.

Thus, the interaction of the physical and social aspects is characterised by choice and, therefore, by the concept of consciousness.

Research Methodology

General Background

The research methodology included an analysis of scientific publications and documents related to the issue of the boundaries between consciousness and self-knowledge in the context of artificial intelligence from a philosophical perspective. The literature review conducted as part of this study was based on materials obtained from such databases as WOS, Web of Science, Scopus, and Google Scholar. It included high quality research and academic papers from various fields such as philosophy, social sciences and engineering. Using the WoSCC system, a bibliometric analysis of the documentation was carried out. The search was carried out in December 2023, and for a highly specialised selection of documents, an advanced search was applied using the Topic Search (TS) option, which allowed identifying the necessary terms in titles, abstracts, authors' keywords and additional keywords (Keywords plus). To ensure broad search coverage, the Boolean operator OR was used to identify documents containing any combination of the requested terms. The analysis included documents (articles, conference proceedings, books, book chapters, review articles, editorials, etc.) for the period from 2019 to 2023. The study considered only scientific articles and review articles, as they were the primary sources of new discoveries, research and critical analysis in the field, and are subject to a thorough peer review provided by scientific journals to achieve established quality standards. WoSCC tools were also used to analyse the data, in particular Analyse Results and Creation Citation Report, and an advanced search using the Topic Search (TS) option, which helped identify terms in titles, abstracts, authors' keywords and additional keywords (Keywords plus). The used keywords were: “artificial intelligence”, “consciousness in AI”, “self-awareness in AI”, “autonomous thinking”, “connectionism”, “philosophy of technology”. To maximise search coverage, we used the Boolean operator OR to identify documents that contained one, two or more of the search terms, as well as an asterisk (*) representing any group of characters, including no characters. For example, the search for “artificial intelligence*” identified documents containing “consciousness in AI”, “self-awareness in AI”, “autonomous thinking”, “connectionism”, “philosophy of technology”, etc.

Research Results

The field of artificial intelligence (AI) was founded in 1956 and is dedicated to the study of intelligent aspects of machine behaviour. The main areas of research in this field can be summarised into two main categories (Fig. 1):

Figure 1

Summarising the main areas of AI research

Source: Davenport (2018).

In the course of the research, despite initial enthusiasm, researchers faced a number of challenges, including the task of creating intelligent models suitable for real-world applications, as noted by Gervais (2021). The main difficulties included a lack of background knowledge, high computational intensity, and limitations of existing knowledge representation systems.

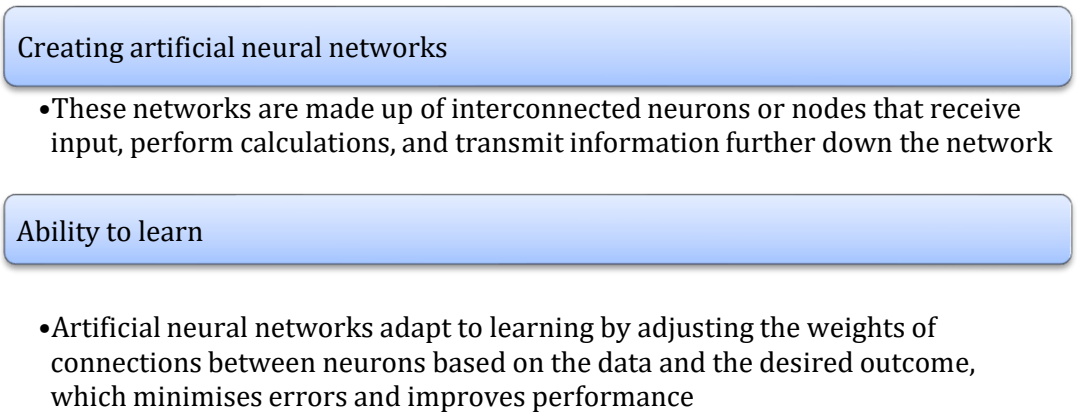
Symbolic artificial intelligence utilises formalised languages like logic to encode knowledge, which is subsequently analysed by cognitive mechanisms. These mechanisms use algorithms to manipulate symbols, allowing for the development of expert and decision support systems capable of drawing conclusions based on established rules and knowledge, as described in Crane (2019).

Unlike other approaches to artificial intelligence, such as machine learning and deep learning, symbolic AI does not require large amounts of data to learn. It relies on knowledge representation and logical reasoning, which makes it particularly useful in areas where knowledge can be clearly defined and represented using logical rules.

Connectionism in the philosophy of artificial intelligence is seen as an attempt to model consciousness and self-awareness by simulating the neural processes that occur in the human brain. According to Da (2019), this approach uses the principles of neural networks to create systems that can process information and learn, similar to biological neurons (Fig. 2).

Figure 2

Key aspects of the connectionist approach



Source: authors' own research.

Thus, the connectionist approach offers a methodology for developing artificial intelligence that can come closer to understanding and reproducing the complex processes of consciousness and self-awareness that are key to human intelligence.

Connectionist systems, which are the basis for understanding consciousness in artificial intelligence, emphasise the importance of parallel data processing. This means that they can perform multiple calculations at the same time, which ensures more efficient and reliable problem solving. The distributed representation of data in connectionist models means that information is distributed among numerous neurons rather than concentrated in one place, allowing the system to adapt to intricate patterns more easily and derive knowledge from just a few examples.

The Deep learning, which is a sub-branch of artificial intelligence, is based on a connectionist approach and focuses on creating complex multi-level neural networks. These models have made significant progress in areas such as computer vision, natural language processing and speech

recognition, demonstrating the ability to automatically learn hierarchical data structures, which increases the efficiency of solving complex problems.

In general, the connectionist approach to artificial intelligence philosophy emphasises the use of artificial neural networks to mimic the cooperative, parallel nature of human brain processing. By learning from data through weight adjustments, connectionist systems have proven to be effective in solving complex problems and achieving impressive results in AI applications.

A neural network, which is a computer analogue of biological neural networks, can serve as a model for the study of consciousness in artificial intelligence. This system, consisting of a mathematical network of artificial neurons organised into layers, allows us to study how the complex phenomenon of consciousness can emerge from simple interactions (Fenves, 2019). Artificial neurons, which are the fundamental elements of a neural network, receive, process, and transmit information, mimicking biological processes. Their connections, influenced by weighting coefficients, adjust over time, similar to the way the brain's plasticity allows for learning .The hierarchical structure of a neural network with input, hidden and output layers allows modelling different levels of information processing, which can be compared to different levels of consciousness. The hidden layers, which transform and combine incoming information, can reflect inner experience, which is a key aspect of consciousness.

Thus, the activation function that each neuron applies to its input signals introduces a nonlinearity that may be important for the generation of conscious experiences. The principle of direct propagation, whereby data passes through layers, reflects the sequence of thoughts that is characteristic of the conscious process (Table 1):

Table 1

The main aspects of neural networks and their learning process

Process	Type of training
Training	A process where neural networks analyse data.
Input data	Data that is fed into the network along with the desired outputs.
Forecasts vs Results	The network compares its forecasts with the desired results.
Correction of scales	The use of algorithms such as gradient descent and back propagation to adjust weights.
An iterative process	A process that allows the network to minimise the gap between forecasts and expected results.
Deep neural networks (DNN)	
Deep learning	

Application. Data processing Complex calculations	Neural networks with several hidden layers. Focuses on training DNNs, learning hierarchical representations and extracting complex patterns. Success in image recognition, natural language processing, and speech synthesis. Ability to process large amounts of data and summarise using examples. Performing complex calculations, making them a powerful tool in AI.

Source: author's own development

This systematisation of AI capabilities helped to structure information about neural networks and their key characteristics, which in turn contributed to a better understanding of their functioning and the possibility of being conscious.

There is general uncertainty about how material objects can have consciousness. This statement reflects the difficulties in understanding the possibility of consciousness in machines. However, such doubts do not force us to reject the idea of creating a machine with artificial consciousness. On the contrary, they stimulate the idea that if we, as material beings, are capable of consciousness, then consciousness may be a material phenomenon that can, in theory, be reproduced artificially.

Consciousness is composed of many diverse and variable elements that cannot be easily separated or explained in isolation. This complexity should be taken into account, because even efforts to explain individual aspects of consciousness are complicated by their interaction. This paper attempts to integrate these different ideas to support the idea of the existence of advanced artificial consciousness in computer systems.

As for artificial consciousness, it is divided into two main types: weak and strong. Weak artificial consciousness imitates conscious actions without understanding the basic principles of consciousness, functioning as a programme that reproduces the behaviour of conscious beings at a detailed level. Strong artificial consciousness, on the other hand, reflects real conscious processes that occur in complex computing systems like artificial brains. The main difference from natural consciousness is the material

that underlies these processes. Also, studies such as Durmishi & Durmishi (2022) point to the existence of numerous intermediate forms of artificial consciousness, which the authors call artificial consciousness shifts.

With the rapid progress in the field of information technology, the relevance of developing a strong artificial consciousness is growing every year. Artificial intelligence, which was once only a topic of science fiction, is now becoming a subject of serious scientific research.

A machine can demonstrate intelligent behaviour by using logic to solve problems because it is programmed to do so. However, there are doubts about its capacity for consciousness, which remain a subject of debate (Gillies & Gillies, 2022). Humans differ from machines in having emotions, experiences, wills, and beliefs. Most of us recognise that our genetics and biology contain certain "programs", but we also believe in our ability to make independent choices that cannot be fully replicated by artificial computer programs that do not possess our unique personal subjective experiences.

This idea would not seem so exciting if there were no prospect of creating a machine capable of producing strong artificial consciousness. Functionalism, a widely cited theory that corresponds to the concept of strong consciousness, defines consciousness through its functions. Although this may seem theoretically simplistic, there are different forms of functionalism. A misunderstanding of qualia does not reflect their ability to be artificially created. It is only part of the broader theory of virtual machine functionalism, which allows that a conscious being can be in many mental states simultaneously, which the virtual machine can account for through its many systems and subsystems.

Rationality is a key criterion for a moral agent and is a basic requirement for the theory of autonomy, which implies the ability to act reasonably. This criterion seems to be the least controversial for artificial agents, so it is mentioned last (Ghoshal & Tucker, 2020). Rationality and logic are characteristics often attributed to artificial agents in popular culture. Modern computers and weak AI systems are known for their ability to perform logical computations, complete complex tasks, and make quick, rational decisions. Nevertheless, the rationality of artificial agents remains a subject of debate.

It should be emphasised that autonomous rationality and general rationality are not synonymous. Autonomous rationality is a cognitive capacity possessed by individuals wherein they are able to engage in decision-making processes based on their rational principles rather than instinctual desires. This capacity allows individuals to exercise their will in a manner that enables them to act in alignment with their personal beliefs and values. It necessitates a deliberate evaluation of potential courses of action before making decisions, and is characterized by a cognitive process that involves both critical thinking and logical reasoning. Through the practice of autonomous rationality, individuals can navigate complex situations and make choices that are consistent with their own rational principles.

In the context of modern computers and weak AI systems, such rationality is not independent. They have no choice and act according to programmed instructions. This is due to the deterministic following of algorithms, which implies the presence of free choice. Virtual machines can be quite complex, capable of phenomenal consciousness, non-determinism (free will), internal purposefulness, and rationality (the ability to experience beliefs, desires, pain, and pleasure). However, despite this, they remain machines. If such a machine reaches a high level of complexity, it will be designed precisely, rationally, algorithmically and architecturally, and the "cold rationality" of algorithmic computers, combined with consciousness, will give it autonomous rationality. Its emotions, phenomenal consciousness, ability to experience pain and pleasure, as well as beliefs and desires, will allow it to overcome hedonistic impulses and make rational and autonomous decisions.

Artificial Intelligence (AI) remains a dynamic field of research that was founded in the 1950s and continues to evolve. Throughout its development, AI has encouraged the rivalry among various research methodologies, sparking the development of fresh challenges and innovative concepts. Despite considerable resistance to theoretical developments, technical progress in this field has brought outstanding results that are rare in the history of science.

Discussion

The main goal of artificial intelligence and its technical applications is to create machines that can mimic the human mind. This task involves research into both physical and mental aspects, which makes it particularly challenging. The wide range of intelligence factors present a complex journey for AI development, with many obstacles stemming from underlying philosophical questions. Many AI specialists have a strong interest in philosophy, and at the same time, advances in artificial intelligence are attracting the attention of the philosophical community.

There is a general uncertainty about how material objects can have consciousness. This statement reflects the difficulties in understanding the possibility of consciousness in machines (Gillies & Gillies, 2022). However, such doubts do not make me reject the idea of creating a machine with artificial consciousness. On the contrary, these findings suggest that consciousness may be a material phenomenon that can potentially be replicated artificially, as evidenced by the idea that our consciousness as material beings could potentially be recreated through artificial means. In general, artificial consciousness is divided into two parts: weak and strong. Weak artificial consciousness is a "simulation of conscious behaviour". It can be implemented as an intelligent program that simulates the behaviour of a conscious being at a certain level of detail without understanding the mechanisms that give rise to consciousness. A strong AC is "a real conscious thought that comes from a sophisticated computing machine (artificial brain)". In this case, the main difference from its natural equivalent depends on the material that generates the process. However, some researchers, such as Durmishi & Durmishi (2022), argue that there are many intermediate areas of consciousness, which they call artificial consciousness shifts.

The consciousness consists of many diverse and variable elements that cannot be easily separated or explained in isolation (Juliani, Arulkumaran, Sasai & Kanai, 2022). This complexity is worth considering, as even efforts to explain individual aspects of consciousness are complicated by their interactions.

This paper sought to integrate these different ideas to support the idea of the existence of advanced artificial consciousness in computer systems.

With the rapid progress in the field of information technology, the relevance of developing a strong artificial consciousness is growing every year. Artificial intelligence, which was once only a topic of science fiction, is now becoming a subject of serious scientific research.

A machine can demonstrate intelligent behaviour by using logic to solve problems because it is programmed to do so. However, there are doubts about its capacity for consciousness, which remain a subject of debate. Humans differ from machines in having emotions, experience, will, and beliefs. While it is true that our genetics and biology dictate certain aspects of our behavior and choices, there is also a significant component of free will that is inherent in human decision-making. This free will is a product of our complex neural networks, which are shaped by both genetic factors and environmental influences. It is this subjective experience that sets us apart from artificial computer programs, which operate based on predetermined algorithms and lack the capacity for true autonomy and creativity. Therefore, while our genetics and biology do play a role in shaping our behaviour, the unique

combination of personal experiences and subjective consciousness enables us to make choices that are not entirely deterministic or replicable by artificial systems. The basic criterion for a moral agent (and the ultimate requirement for the theory of the three conditions of autonomy, i.e., acting reasonably) is rationality. This criterion is perhaps the least controversial for an artificial agent, so I put it last. Rationality and logic are the defining characteristics of artificial agents in popular culture. Modern computers and weak AI systems are notorious for their logic. They perform large computations and can make extremely complex decisions quickly and rationally. However, the rationality of artificial agents is not without controversy.

This idea would not seem so exciting if there were no prospect of creating a machine capable of producing strong artificial consciousness. Functionalism, a widely cited theory that corresponds to the concept of strong consciousness, defines consciousness through its functions. Although this may seem theoretically simplistic, there are different forms of functionalism. A misunderstanding of qualia does not reflect their ability to be artificially created. It is only part of the broader theory of virtual machine functionalism, which allows that a conscious being can be in many mental states simultaneously, which the virtual machine can account for through its many systems and subsystems.

Rationality is a key criterion for a moral agent and is a basic requirement for the theory of autonomy, which implies the ability to act reasonably. This criterion seems to be the least controversial for artificial agents, so it is mentioned last. Rationality and logic are characteristics often attributed to artificial agents in popular culture. While modern computers and weak AI systems are renowned for their prowess in executing logical computations, tackling intricate tasks, and delivering speedy, rational decisions, the rationality of artificial agents continues to be a topic of discussion. Artificial Intelligence (AI) remains a dynamic field of research that was founded in the 1950s and continues to evolve. Throughout its development, AI has fostered competition between different research methods, leading to the emergence of new problems and ideas. Despite considerable resistance to theoretical developments, technical progress in this field has brought outstanding results that are rare in the history of science.

The main goal of artificial intelligence and its technical applications, according to Pak, Hurbanska, Boiko, Avramenko & Katerynych (2023), is to create machines that can mimic the human mind. This task includes research into both physical and mental aspects, which makes it particularly challenging. The diversity of aspects of intelligence leads to a challenging path of development in the field of AI, and many of the challenges faced by the field are closely related to philosophical issues. Many AI professionals have a strong interest in philosophy, and at the same time, advances in AI are attracting the attention of the philosophical community (Lee & Lee, 2020).

The main task of cognitive science in the context of artificial intelligence is to gain a deep understanding of the anatomy and mechanisms of human consciousness, as well as to develop a sound explanation of the interaction between intelligence, emotions, and intentions within human consciousness. AI researchers are working to formalise these processes of consciousness (Reggia, Katz & Davis, 2020). In order to recreate human consciousness, it is necessary to first understand its structure and dynamics. The question of whether AI can possess consciousness depends on determining its actual presence. According to a study by Glikson & Woolley (2020), the cognitive sciences are driving advances in artificial intelligence, which is a key factor in the cognitive turnaround. This process involves the interaction between philosophy, cognitive psychology, neuroscience, brain science, artificial intelligence, and other areas that are interconnected regardless of the areas of computer science development, from symbolic systems to expert systems, knowledge engineering, biological computers, and quantum computing.

In a related study, Harris, Nambiar, Rajasekharan & Gupta (2020) noted that this issue was closely related to a deep understanding of human consciousness and its aspects through the lens of philosophy. Regardless of whether the approach belongs to a strong or weak school of artificial intelligence, from an epistemological point of view, artificial intelligence is based on symbolic systems to mimic some aspects of human thinking. However, the accurate reproduction of human consciousness depends not only on technological innovations, but also on the philosophical understanding of consciousness and the factors that influence it. In the modern context, the philosophical question of artificial intelligence is not so much about its essence as about solving specific problems of intellectual modelling (Yu, 2019).

To answer these questions, Jalilbayli considers the concept of intentionality as the ability to anticipate and plan actions, which is a fundamental aspect of consciousness. In the realm of AI, intentionality refers to a system's capacity to autonomously make decisions using its internal understanding and perception of the environment (Liu, Ouyang, Wang, Fieguth, Chen, Liu & Pietikäinen, 2020). However, whether this ability to make independent decisions can be equivalent to human consciousness remains an open question.

Regarding the potential harm that AI can cause, Jones (2022) emphasises that intentionality does not necessarily include malice or intent to harm. AI generally acts in accordance with its programming instructions, and any harm caused by the system is usually the result of programming errors or unforeseen interactions with complex external conditions. Thus, the question of AI's intentionality and its capacity for consciousness remains a subject of active research and philosophical debate.

In the context of intelligence, there is the question of whether machines are able to use intelligence to solve problems in the same way that humans do, or whether there is a threshold beyond which a machine cannot cross to solve complex problems. Kaliuta (2023) notes that humans can tap into their unconscious abilities that are not available for formalisation. We cannot transfer to computers the knowledge that we ourselves cannot clearly articulate (Maidaniuk, Tetiana, Hoian, Doichyk, Patlaichuk & Stupak, 2022). To get around this problem, computer scientists did not try to copy human intelligence, but instead created a new approach to thinking for artificial intelligence based on data analysis.

In her paper, Maraieva (2022) says that the artificial intelligence conundrum can be considered the newest philosophical paradox. Our incomplete understanding of learning processes in the human brain leads us to allow artificial intelligence to process information statistically. Paradoxically, the knowledge of the mechanisms of artificial intelligence remains limited, creating a situation with two incomprehensible systems (Müller, 2020). This phenomenon is often described as the “black box problem”, where we know what is input to the system and what results it produces, but the decision-making mechanism remains unknown. Due to their lack of linguistic abilities, artificial neural networks cannot explain their actions or motives, and like all artificial intelligence, they do not possess common sense.

A growing concern among the public is the potential bias that may be built into some artificial intelligence (AI) algorithms, including sexism and racial discrimination (Poznanski, Cacha, Sbitnev, Iannella, Parida, Brandas & Achimowicz, 2023). One example is new programmes designed to predict the risk of repeat criminal behaviour, which can be particularly harsh on certain groups. If AI works with reliable data, its conclusions can be accurate, but often they reflect existing human biases (Parmar, 2023). In the realm of ethics, the concern about the possible harm machines could cause to humans is a prominent issue. Researchers are actively seeking methods to guarantee that machines act in an ethical manner to prevent any risks to human safety. As for emotions, scientists are divided on the topic of whether machines are capable of experiencing human emotions like love or hatred (Rafanelli, 2022).

The general consensus is that artificial intelligence should not be expected to consciously strive for good or evil.

Considering potential risks, Shrivastava (2017) identify two main scenarios:

AI designed to perform destructive tasks: for example, autonomous weapons designed to kill could become a tool of great harm if they fall into the hands of malicious actors. An AI arms race could lead to unpredictable consequences, including war, which would cause numerous casualties.

Loss of control over complex "closed" weapon programmes, which can lead to uncontrollable situations. Although this risk is present in the field of specialised AI, it increases with the development of more intelligent systems and autonomy.

According to a study by Sengupta, Garg, Choudhury, and Aggarwal (2018), artificial intelligence (AI) was created to perform useful tasks, but sometimes the processes it performs can have negative consequences. This is especially true when the goals of humans and AI do not coincide, which is a difficult problem to solve.

Suter (2019) points out the problems that arise from the conceptual foundations of AI. Science is based on what is already known, and the ability to make scientific observations is limited by that knowledge. We use the known to understand the unknown, and the two aspects always exist in interdependence (Skansi & Kardum, 2021). Without the known, it is impossible to explore the unknown, and without the unknown, science cannot develop (Van Heerden & Bas, 2021). The human experience of objects is not limited to the light entering the eye, but also depends on many other factors that shape perception.

The theories relating to consciousness in the context of artificial intelligence (AI) are universal and complex, which highlights the need for a theoretical framework for observation (Wang, Zhao & Pourpanah, 2020). Scientific research requires theory as a basis for understanding, and this understanding cannot be based on the unknown. Businesses often face difficulties in determining optimal strategies, so AI-enabled consulting services try to guide business processes. At the same time, the influence of philosophy on AI research and application has its weaknesses that require further analysis.

Conclusions and Implications

Although technology and philosophy have different methods, both areas focus on the same subject - the human being. The technological progress is aimed at solving certain practical problems that arise in everyday life in order to improve people's lives in the near future. However, technological innovations are usually limited to solving these specific problems and do not go beyond them, as this is not necessary for a technical solution to the problem. The main goal of technology is to be useful, which reflects its instrumental nature.

Philosophy goes beyond surface-level concerns and practical matters to delve into the deeper complexities of human existence. It seeks to gain a comprehensive understanding of a subject by exploring not just its surface characteristics, but also its ethical implications and influence on broader human concerns. This includes analysing the origin, evolution and essence of values. Thus, philosophy focuses on in-depth analysis and criticism of general ideas and contemporary events to identify changes in value systems. On the other hand, technologies, especially artificial intelligence, are focused on solving specific existing problems, which may seem paradoxical compared to the philosophical approach.

This undoubtedly contributes to the advanced development of the latest technologies. A design that foresees many possible outcomes, just like the technical solution itself, allows future problems to be solved reliably. This is especially true of artificial intelligence, which can now be defined as "the science and technology of creating intelligent machines, including intelligent software". Philosophers have contributed to the development of the key concepts of AI, including defining the characteristics an artefact needs to be recognised as intelligent and the basic principles of rationality that also derive from philosophical discourse, what is even more exciting is the role of philosophy in guiding the development of AI and its integration into human lives as it examines the broader context of its impact.

Suggestions for Future Research

Future research could investigate phenomenological approaches, such as those used by Husserl and Merleau-Ponty, to understand how AI's potential self-awareness aligns with modern philosophies of consciousness. This exploration may help simulate the subjective experience of modern philosophy. The study of the ethical implications of AI capabilities and the attribution of consciousness or self-awareness to AI is a promising area. The discussion of the potential implications for AI development, human interaction with AI, and societal norms regarding artificial entities can help to understand and define ethical frameworks and regulations in this area.

A comparative analysis of AI with animal consciousness is also promising. Comparing AI's assumed self-awareness with that of non-human animals and assessing similarities and differences in cognitive processes, behavioural responses, and ethical considerations regarding rights and moral status is a new challenge for contemporary philosophical thought.

The integration of Eastern philosophical traditions is appropriate. Exploring the views of Eastern philosophies, such as Buddhism or Taoism, on consciousness and self-awareness and considering how these approaches can enrich philosophical understanding of AI's potential cognitive abilities could influence the results of current related research.

The study of the impact on personal identity and social relations can be innovative to the traditional understanding of personal identity and human relations. The consideration of philosophical concepts of identity and the role of consciousness in shaping interpersonal dynamics and inclusion in metaphysical research on the nature of consciousness itself and whether, being self-aware, AI has a fundamentally different consciousness from human consciousness is a potentially promising area of work.

The practical aspects of AI development in terms of continuing and expanding the presented work may contribute to the study of practical methodologies for verifying and confirming the claims of AI self-awareness. Furthermore, shedding light on this issue will pave the way for establishing a set of ethical guidelines and regulations that factor in the cognitive capacities of AI. This promising research aims to deepen the philosophical investigation of AI's cognitive abilities and to consider the broader ethical, metaphysical, and societal implications.

Acknowledgements

None.

Conflict of Interest

None.

Funding

The Author received no funding for this research.

References

Agarwal, A., & Edelman, S. (2020). Functionally effective conscious AI without suffering. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 39-50. https://doi.org/10.1142/S2705078520300030

Assunção, G., Patrão, B., Castelo-Branco, M., & Menezes, P. (2022). An overview of emotion in artificial intelligence. *IEEE Transactions on Artificial Intelligence*, *3*(6), 867-886. https://doi.org/10.1109/TAI.2022.3159614

Bajohr, H. (2020). *Algorithmic empathy: On two paradigms of digital generative literature and the need for a critique of AI works*. doi:10.5451/UNIBAS-EP79106

Barron, L. (2023). AI and Literature. In *AI and Popular Culture* (pp. 47–87). Emerald Publishing Limited. https://doi.org/10.1108/978-1-80382-327-020231003

Brinkman, D., & Grudin, J. (2023). Learning from a Generative AI Predecessor--The Many Motivations for Interacting with Conversational Agents. https://doi.org/10.48550/arXiv.2401.02978

Chun, J., & Elkins, K. (2022). What the Rise of AI Means for Narrative Studies: A Response to "Why Computers Will Never Read (or Write) Literature" by Angus Fletcher. *Narrative*, *30*(1), 104-113. https://doi.org/10.1353/nar.2022.0005

Crane, G. (2019). Beyond Translation: Language Hacking and Philology. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.282ad764

Craswell, N., Mitra, B., Yilmaz, E., Campos, D., & Voorhees, E. M. (2020). Overview of the TREC 2019 deep learning track. https://doi.org/10.48550/arXiv.2003.07820

Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry*, *45*(3), 601–639. https://doi.org/10.1086/702594

Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73-80. https://doi.org/10.1080/2573234X.2018.1543535

Dodda, S. B., Maruthi, S., Yellu, R. R., Thuniki, P., & Reddy, S. R. B. (2021). Ethical Deliberations in the Nexus of Artificial Intelligence and Moral Philosophy. *Journal of Artificial Intelligence Research and Applications*, *1*(1), 31-43. https://aimlstudies.co.uk/index.php/jaira/article/view/25.

Durmishi, L., & Durmishi, A. (2022). A philosophical assessment of social networks impact on adolescents' development in conditions of unlimited access to information. *Future Philosophy*, *1*(2), 27-41. https://doi.org/10.57125/FP.2022.06.30.03

Edelman, S. (2023). On the ethics of constructing conscious AI. In *Computational Approaches to Conscious Artificial Intelligence (*pp. 273-294). https://doi.org/10.1142/9789811276675_0010

Esmaeilzadeh, H., & Vaezi, R. (2021). Conscious AIhttps://doi.org/10.48550/arXiv.2105.07879

Fenves, P. (2019). "Einstein's Brain" in Three Parts. *The Yearbook of Comparative Literature*, *62*(1), 174-188. https://doi.org/10.3138/ycl.62.005

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, *30*(3), 411-437. https://link.springer.com/article/10.1007/s11023-020-09539-2

Gervais, D. J. (2021). AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines. *Seton Hall L. Rev.*, *52*, 1111. https://heinonline.org/HOL/LandingPage?handle=hein.journals/shlr52&div=37&id=&page=

Ghoshal, B., & Tucker, A. (2020). Estimating uncertainty and interpretability in Deep Learning for Coronavirus (COVID-19) detection. In *arXiv [eess.IV]*. http://arxiv.org/abs/2003.10769

Gillies, D., & Gillies, M. (2022). Artificial Intelligence and Philosophy of Science from the 1990s to 2020. In *Current Trends in Philosophy of Science: A Prospective for the Near Future* (pp. 65-79). Cham:

Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-01315-7_4

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: A review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Haikonen, P. O. (2020). On artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 73-82. https://doi.org/10.1142/S2705078520500046

Harris, P., Nambiar, R., Rajasekharan, A., & Gupta, B. (2020). AI Powered Analytics App for Visualising Accident-Prone Areas. In *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2018* (pp. 361-367). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-19562-5_36

Heflin, J. J. A. (2020). *AI-generated literature and the vectorised Word* (Doctoral dissertation, Massachusetts Institute of Technology). https://dspace.mit.edu/handle/1721.1/127563

Jalilbayli, O. B. (2022). Philosophy of linguistic culture and new perspectives in modern azerbaijani linguistics. *Future Philosophy*, *1*(4), 53-65. https://doi.org/10.57125/FP.2022.12.30.05

Jones, N. (2022). Experiential Literature? Comparing the Work of AI and Human Authors. *APRIA Journal*, *5*(5), 41-57. https://doi.org/10.37198/APRIA.04.05.a5

Juliani, A., Arulkumaran, K., Sasai, S., & Kanai, R. (2022). *On the link between conscious function and general intelligence in humans and machines*. https://doi.org/10.48550/ARXIV.2204.05133

Kaliuta, K. (2023). Integration of AI for Routine Tasks Using Salesforce. *Asian Journal of Research in Computer Science*, *16*(3), 119-127. https://doi.org/10.9734/ajrcos/2023/v16i3350

Kampen, K., Romanchuk, S., & Palij, D. (2022). Confessional style of the Ukrainian language. *Occasional Papers on Religion in Eastern Europe*, *42*(2), 11. https://doi.org/10.55221/2693-2148.2331

Lee, R. S. T. (2020). AI and self-consciousness. In *Artificial Intelligence in Daily Life* (pp. 349–368). Springer Singapore. https://link.springer.com/chapter/10.1007/978-981-15-7695-9_13

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, *128*(2), 261–318. https://doi.org/10.1007/s11263-019-01247-4

Maidaniuk, I., Tetiana, T. S. O. I., Hoian, I., Doichyk, M., Patlaichuk, O., & Stupak, O. (2022). The problem of artificial intelligence in contemporary philosophy. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *13*(4), 436-449. https://edusoft.ro/brain/index.php/brain/article/view/1348.

Maraieva, U. (2022). On the formation of a new information worldview of the future (literature review). *Future Philosophy*, *1*(1), 18-29. https://doi.org/10.57125/FP.2022.03.30.02

Müller, T. (2020). Conscious, thinking, and intelligent machines?: Some clarifying remarks in the field of articial intelligence. In *Artificial Intelligence* (pp. 3–14). Brill | mentis. https://doi.org/10.30965/9783957437488_003

Ng, G. W., & Leung, W. C. (2020). Strong artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 63-72. https://doi.org/10.1142/S2705078520300042

Nikolenko, K. (2022). Artificial Intelligence and Society: Pros and Cons of the Present, Future Prospects. *Futurity Philosophy*, *1*(2), 54–67. https://doi.org/10.57125/FP.2022.06.30.05

Pak, A., Hurbanska, S., Boiko, O., Avramenko, V., & Katerynych, P. (2023). The formalized semantics of neologisms-slangisms in the context of the English translation of A military narrative. *World Journal of English Language*, *13*(6), 537. https://doi.org/10.5430/wjel.v13n6p537

Parmar, D. (2023). *Enhancing customer relationship management with Salesforce Einstein GPT* [Bachelor's thesis, Haaga-Helia University of Applied Sciences]. Theseus. https://www.theseus.fi/bitstream/handle/10024/812434/Parmar_Dipal.pdf?sequence=2&isAllowed=y

Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, *122*(106855), 106855. https://doi.org/10.1016/j.chb.2021.106855

Poznanski, R. R., Cacha, L. A., Sbitnev, V. I., Iannella, N., Parida, S., Brandas, E. J., & Achimowicz, J. Z. (2023). Intentionality for better communication in minimally conscious AI design. *Journal of Multiscale Neuroscience*, *3*(1), 1-12. https://doi.org/10.56280/1600750890

Rafanelli, L. M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society*, *9*(1), 205395172210806. https://doi.org/10.1177/20539517221080676

Reggia, J. A., Katz, G. E., & Davis, G. P. (2020). Artificial conscious intelligence. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 95-107. https://doi.org/10.1142/S270507852050006X

Sengupta, E., Garg, D., Choudhury, T., & Aggarwal, A. (2018, November). Techniques to eliminate human bias in machine learning. In *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 226-230). IEEE. https://doi.org/10.1109/SYSMART.2018.8746946

Shrivastava, M. (2017). *Learning Salesforce Einstein*. Packt Publishing Ltd. https://www.packtpub.com/product/learning-salesforce-einstein/9781787126893.

Skansi, S., & Kardum, M. (2021). A Prolegomenon on the Philosophical Foundations of Deep Learning as Theory of (Artificial) Intelligence. *Disputatio philosophica: international journal on philosophy and religion*, *23*(1), 89-99. https://hrcak.srce.hr/ojs/index.php/disputatio-philosophica/article/view/22252

Suter, R. (2019). Artificial intelligence and the cloud. In *Artificial Intelligence and Machine Learning for Business for Non-Engineers* (p. 27). Retrieved from https://books.google.com.ua/books?hl=uk&lr=&id=-BTADwAAQBAJ&oi=fnd&pg=PA27&dq=Salesforce+Einstein+AI++linguistics&ots=9NESpYqW2a&sig=R_OccSawBiJmvpPgA_ycXh6HMgs&redir_esc=y#v=onepage&q&f=false

Van Heerden, I., & Bas, A. (2021). Viewpoint: AI as author – bridging the gap between machine learning and literary theory. *The Journal of Artificial Intelligence Research*, *71*, 175–189. https://doi.org/10.1613/jair.1.12593

Wang, X., Zhao, Y., & Pourpanah, F. (2020). Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, *11*, 747-750. https://link.springer.com/article/10.1007/s13042-020-01096-5

Wei, C. (2022). Copyright protection and data reliability of AI-written literary creations in smart city. *Security and Communication Networks*, *2022*, 1–13. https://doi.org/10.1155/2022/6498468

Yu, J. (2019). *Getting started with Salesforce Einstein analytics: A Beginner's guide to building interactive dashboards*. Apress. https://link.springer.com/10.1007/978-1-4842-5200-0

Zhao, G., Li, Y., & Xu, Q. (2022). From Emotion AI to cognitive AI. *International Journal of Network Dynamics and Intelligence*, 65–72. https://doi.org/10.53941/ijndi0101006